

Upheaval for British research?

After nearly a decade of stagnation, schemes for reorganizing the British research councils are in the air again. That is as it should be.

THE British research councils, cruelly and short-sightedly deprived for the past eight years of the funds needed properly to fulfil their functions, have nevertheless taken their own continued existence for granted throughout that time. It is as if being short of funds, however seriously, is one of the unavoidable hazards of the public life, but that the prospect of being reorganized, with the threat it carries for continued existence in any form, is too great to be openly entertained. So much is understandable. The puzzle is that the four British research councils concerned with the natural sciences, having stuck to the doctrine of immutability all this time, should now have taken to babbling about their constitution in the unlikely setting of the evidence they have provided to a subcommittee of the House of Lords Select Committee on Science and Technology, which is embarked on a study of agricultural research in Britain (see *Nature* 333, 109; 1988). One council (the Natural Environment Research Council, NERC) wants to merge with another (the Agriculture and Food Research Council, AFRC), one (the Medical Research Council, MRC) wants biological research taken away from another (the Science and Engineering Research Council, SERC), which rejects the idea. If the councils carry on like this, they must expect to be taken seriously — and reorganized.

Not before time. Over the past few years, the case for change has been growing, for reasons which are structural. NERC is manifestly the least well placed to continue in its present form. The failure, nearly 30 years ago, to detach the Meteorological Office from the British Ministry of Defence lop-sided NERC's field of interest, from the beginning an uncomfortably large proportion of its funds has been committed to in-house research institutes and, more recently, the freedom of manoeuvre derived from research contracts with government ministries has been constrained as the ministries have turned away from research. NERC could not in these circumstances provide the funds adequately to support academic research in, say, geophysics. NERC's case for being part of a larger organization is overwhelming; its case for merging with AFRC hangs on the £19 million a year it spends on biological research (17 per cent of the total), which is less convincing.

Biology

MRC's case for robbing SERC of its interest in biology research (costing £23 million a year) is different. The council says that its own spending on non-clinical research in universities is more than that (£27 million a year), that AFRC is also active in the field, that coordination of these programmes has been more difficult than the (smaller) task of coordinating the human nutrition research supported by both MRC and AFRC and that, in any case, support for biology research at British universities should be channelled through "mission-led" research councils (which number NERC, AFRC and MRC). MRC says it wants to be given more responsibility for "exploitable biological research", that it recognizes this would entail liaison with industry and that it has the management to do the job. The logic of the argument leaves much to be desired, not least the implication that all biology should be exploitable.

Poor SERC, the largest of the research councils, is on this occasion the most defensive. Its case for the *status quo* (but "there may be some argument" for merging NERC's biology with AFRC's) is its record, a list of useful biology projects on which both SERC and industrial companies have supported collaborative research in universities. So far as it goes, the case is telling, but the truth about SERC is that the bodies are buried elsewhere. Over the years, the provision of central facilities for the support of academic research has grown so large that, with straitened budgets, it has become the tail that wags the dog. SERC's handling of subscriptions to international facilities is a constant embarrassment. And SERC, being large and therefore conspicuous, is too readily partly hijacked by government departments bent on new schemes for supporting British industry. Curiously, in their stream-of-consciousness statements in the House of Lords, none of the councils raised the forbidden question of what will happen to the £600-million-plus spent by the University Grants Committee each year, ostensibly in support of academic research.

Reorganization

Where all this will lead is anybody's guess, simple though it is to tell where it should lead. The time has indeed arrived for radical reorganization of the British research council system, but only in the light of a better general understanding of its purpose. Support of academic research by means of competitively awarded research grants, in which all the councils now have a hand, is one central function. Whether research councils should themselves support graduate students (and even allocate them to university departments) is another more open question. All the four councils also support much applied research, either because that is their chief remit (MRC and AFRC) or at the behest of government (mostly affecting NERC and SERC). In other roles, they act as middlemen for research (using government department funds to commission academic research), as research contractors in their own right and research entrepreneurs. It is no wonder that research council heads are harassed people; luckily, they may think, it will not be for them, but for the Advisory Board for the Research Councils (ABRC) eventually to suggest how the impending reorganization should be shaped. But that may be an illusion. ABRC itself cannot be exempted from the changes ahead.

This is how the job should be done. The British research enterprise, weakened as it has been by the past decade, needs a central body to battle with the British government over matters of policy and otherwise to determine priorities between very different fields of research. That should be a renamed ABRC, which should also be given responsibility for external subscriptions (and for deciding which to continue to pay) as well as for general matters such as the support of graduate students. There should be a predominant source of academic research grants, SERC without the trappings, which should be encouraged (compelled?) to hive off to university consortia its central facilities not so much for budgetary reasons as to avoid conflicts of interest (as in the suspicion that research grants keeping central facilities busy have an inside track). It would be ridiculous that



biology and molecular biology in particular should be separated from this mechanism, but MRC (dwarfed as it is by the recent spectacular growth of private charitable research funds) would still have an important function in fields such as human genetics. But direct academic support (now small) by NERC and AFRC should be transferred to the new SERC, and should themselves be merged into an applied research council (together with SERC's own interests in industrial projects, as in biotechnology and electronics). And an attempt should be made, perhaps by means of an applied research council, to breathe new life into the best of the Rothschild recommendations in 1971 — that departments of the British government should be made competent in knowing what research could do for them.

The benefits of some such reorganization would be several, not the least that it would be easier for the several research councils to tell what they are for. Only in that way is British academic research likely to win the benefits of stability of which it has been too much deprived in recent years. The need for acting quickly is all the more necessary now that the British university system has been set off on the long and painful road towards the concentration of research resources on chosen institutions and departments. Especially with that prospect, it is unseemly that grown men should be quarrelling in the House of Lords about questions such as whether photosynthesis belongs more properly with AFRC or SERC; they should instead be wondering about the restoration of the British research enterprise to good health. But a radical change along these lines would have the further advantage that the armies of selfless men and women who at present, by their membership of research councils and their committees, are all perpetually brooding about the same vast range of intractable problems would have a more comprehensible set of issues on their plates. □

Second term, second try

Will the same recipe for research serve M. Mitterrand as well during his second presidency?

THE return of President François Mitterrand in last Sunday's election will not restore the mood of French science to that just eight years ago, when Mitterrand's first research minister, Jean-Pierre Chevènement, set about throwing money at almost every problem in sight. Not only is M. Hubert Curien, Chevènement's successor (with a much less grand portfolio of interests) a naturally more cautious man, but Mitterrand himself is also wiser as well as older. Chevènement was asked to carry the torch for research at a time when the president himself believed that France was ready for a thoroughgoing reform of its social and economic institutions. In the event, the reform came to a halt when France's bankers took fright at the government's deficit, and when the voters who re-elected President Mitterrand on Sunday gave him an assembly dominated by his opponents. Mitterrand would not relish a re-run of that course.

Nor need he take the risk. The most curious feature of the past eight years in France is that, over-optimism and false starts notwithstanding, so much has been done to change the climate for science and technology. While the roots of this process go back to the gaullist era at the end of the 1950s, Mitterrand's first spell in office as the first socialist president of this republic has brought a rich harvest of technical progress and, more important, a dramatic deepening of the French research enterprise at university and research institute laboratories. French research and French industry are in good shape, and it has ceased to be a marvel that the telephone system works. The worm in the apple is merely that France remains prone to the seductions of flashy schemes, among which its devotion to man-in-space is only the most conspicuous. Task-directed programmes may have served well in the first of Mitterrand's presidencies, but Curien would be well-advised, on his behalf, to watch out for their dangers on this second round. □

More embryo research?

The British committee overseeing the practice of IVF may not have much life ahead of it.

THIS time next year, the British government will almost certainly have carried legislation onto its statute book to regulate research with fertilized human eggs. The government has been explicit about its policy (published last November, see *Nature* 330, 407; 1987) — clinics and practitioners that practise *in vitro* fertilization will have to be licensed, commercial surrogacy will become illegal and so on, but the question of whether research with fertilized human eggs will be allowed will be determined by the British parliament. On that part of the government's intended bill, to be published in the autumn, Members of Parliament will not be dragooned or cajoled into one or other of the voting lobbies, but will instead be allowed to decide for themselves.

This is the sense in which last week's third report of the Voluntary Licensing Authority (VLA) for *in vitro* fertilization (IVF) and embryology may be the last-but-one of its kind. The committee owes its existence to the correct estimation of the Medical Research Council and the Royal College of Obstetricians and Gynaecologists that, in the absence of legislation, self-regulation would be prudent. But the committee has no statutory backing, but must rely on practitioners and researchers to collaborate by requesting licences to do what they seek. Those concerned are sufficiently aware of the benefits research will derive from public trust that there is not a scintilla of doubt that voluntary applications have been comprehensive. The fear now is that that state of grace may not continue.

Yet the research recorded in this third report (a total of 18 projects voluntarily notified in 1987) is also a vivid proof of how innocuous are people's ambitions. Almost all the projects from 1987 and earlier years are concerned with the technology of IVF itself — what are the determinants of success in IVF procedures and how might success-rates be improved? There are just two (of 18) projects in which people are using genetic probes to look for genetic defects in fertilized eggs. This is hardly the Frankenstein stuff of which people have been worrying.

The report provides telling evidence both of the demand for IVF in Britain and of the still poor rate of success even with present techniques. That there should have been 4,670 patients at IVF clinics in 1986 is one measure of the demand for the service. That a large number of women should have endured a total of more than 7,000 oestrus cycles after being implanted with fertilized eggs in the hope of becoming pregnant is a measure of the sacrifices people are prepared to make to relieve infertility. That there were only 605 live births from all this effort is a proof that IVF remains a potent source of high hope disappointed. The report also carries telling evidence of how further research is almost certain to improve success. At six of the larger British centres, responsible for treating two-thirds of all infertile patients, nearly a fifth of all egg transfers led to a live birth, more than six times the success rate at the smaller clinics. IVF is plainly a technique in great demand which is much in need of improvement.

That is the context in which the British parliament (and others elsewhere) should regard the issue it will face later in the year. The crying need is for a substantial improvement of the technique. To prohibit investigations leading to improvement would be to perpetuate the present state of affairs in which thousands of women each year take pot-luck with a new technique, and are cruelly disappointed four times out of five. The VLA itself, having produced the evidence, "strongly recommends" that the liberal option of notified research with strict safeguards should be applied to embryo research, but prudently, it rests its case at that. But this declaration is set in such a detailed and constructive appraisal of the government's proposals that it will require great obstinacy by parliamentarians to reject it. Unfortunately, that is no assurance that they will not. □

Candidate cause identified of non-A, non-B hepatitis

- Elusive virus isolated after long search
- Potential market for blood screening

Washington

THE search for the elusive viral agent responsible for non-A, non-B hepatitis may be over. Last week, the California biotechnology company Chiron announced, to an accompaniment of stock trading activity, that its researchers had isolated and partially characterized the virus responsible for a large proportion of cases. The commercial prize is the sale of blood-screening diagnostic tests in the near term and of vaccines in the future.

The cause of blood-borne hepatitis cases not attributable to hepatitis-B virus has been sought for more than 15 years. (Hepatitis A, or infectious hepatitis, is

transmitted through contaminated food or water.) Although tests for the blood-borne hepatitis B were developed in the early 1970s, until 1986 up to 10 per cent of transfusion recipients were still contracting hepatitis, presumably from infection by an unknown agent.

Because the virus, or viruses, responsible had not been identified, no specific test to screen transfusion blood has been available. Instead, since 1986 blood banks have used two surrogate tests, one for the presence of antibodies to hepatitis-B core antigen, the other for elevated levels of alanine aminotransferase. Positive results on both tests correlate with

risk for infection with non-A, non-B hepatitis.

The surrogate screening reduced the incidence of non-A, non-B hepatitis by half, but roughly 200,000 people each year who receive blood transfusions in the United States are still exposed to the non-A, non-B hepatitis virus. Of these, half may go on to develop chronic hepatitis, and 20 per cent of the chronic carriers will eventually develop cirrhosis.

Chiron says it has been working on the problem for the past five years. Data suggesting that Chiron researchers may have isolated the virus responsible for the majority of non-A, non-B hepatitis cases were reported at a small symposium at the University of California at San Francisco on Monday last week. Chiron said that because of the commercial significance of the discovery, it held press conferences the next day to allow all its shareholders to have equal access to the information.

According to Michael Houghton, Chiron has cloned and expressed a protein from a virus that reacts with antibodies from over half of patients who have non-A, non-B hepatitis. Houghton says the virus is a single-stranded RNA virus with a genome of approximately 10,000 nucleotides, which may be classed as a toga virus.

The virus has been difficult to study because it is found in very low concentrations in infected tissue, and only infects chimpanzees and humans. But Houghton's group has sequenced 30–40 per cent of the virus, and has expressed roughly 10 per cent of its coding potential through recombinant means. Houghton hopes to have his group's results published later this year.

The announcement last week spurred sales of Chiron stock, and drove the price per share up by 30 per cent at one point. Officials of the NASDAQ over-the-counter market froze trading of the stock temporarily on Tuesday last week because of the large volume of shares changing hands, suggesting that word had leaked out before the press conferences began.

Chiron has developed a prototype immunoassay for the new virus, and plans to apply for human trials before the end of 1988. The potential market for the diagnostic test in blood-screening and clinical diagnosis is estimated at \$85 million a year. Chiron has licensed marketing rights for the immunoassay to Ortho Diagnostic Systems, a subsidiary of Johnson & Johnson.

A vaccine for non-A, non-B hepatitis is still far off, but Chiron has formed (with Ciba-Geigy) the Biocine Corporation to market any that is developed. Chiron produced the first recombinant vaccine, for hepatitis B, now sold by Merck. Chiron has also applied for patents covering the expression of non-A, non-B hepatitis virus proteins.

Carol Ezzell

Conflicting views on the safety of US biological weapons

Boston

CONTROVERSY over the safety of US biological weapons programmes continued last week with the release of conflicting reports by the Pentagon and by the majority staff of a Senate oversight subcommittee.

The Pentagon assessment, which says the programmes do not imply adverse environmental or public health effects, stems from a 1986 lawsuit brought by Jeremy Rifkin's Foundation on Economic Trends. On the grounds that the military's conduct of its biological defence research involved "wholesale violations of the

National Environmental Policy Act," Rifkin's group last year won a court settlement requiring the Pentagon to write an environmental impact statement, of which last week's report is the draft.

Citing many "mitigative measures, controls and monitoring", the Pentagon says its biological research shows that the perceived risks of the research programme are "very much exaggerated" and that even the unlikely event of a severe accident would not lead to "catastrophic results".

But Andrew Kimbrell, a lawyer for the Foundation for Economic Trends, called the Pentagon assessment "shoddy", "irresponsible" and "totally inadequate". He says that unless the report is substantially revised, the foundation will return to court seeking an injunction of the entire biological research programme.

The Senate report, timed to coincide with the Pentagon assessment, also differs sharply with it, stating that there are "increasing risks" and "serious failings" in the Defense Department's management of safety controls. These are said to include "inadequate regulations, lax safety enforcement and documented safety lapses".

Democratic Senator Carl Levin, chairman of the Senate Subcommittee on Oversight of Government Management, says that his subcommittee will hold hearings this summer on problems raised in the report and that it will also carry out its own review of the draft environmental impact statement.

Seth Shulman

Fang Lizhi returns

London

DR Fang Lizhi, the astrophysicist who in January 1987 was dismissed from his post as vice chairman of China's University of Science and Technology and expelled from the Communist Party, has made an unofficial political comeback, addressing a gathering of students on the topics of human rights, democracy, economic reform and modernization.

Fang's political disgrace followed an address he gave to university students at the end of 1986, which, the party authorities alleged, had incited them to street demonstrations demanding greater intellectual freedom. On this occasion, Fang made it clear that he was not in favour of the students taking to the streets, but rather they should "keep quiet for a time".

Vera Rich

Pasteur takes US company to court

Paris

THE Pasteur Institute, which holds the patent rights for the test kit used to detect the second human immunodeficiency virus (HIV-2) known to cause AIDS, and Diagnostics Pasteur, which holds the exclusive licence to these rights, has accused the US company, DuPont de Nemours Inc., as well as its French, Swiss and Belgian subsidiaries, of patent infringement.

HIV-2 was first isolated by researchers at the Pasteur Institute in 1985 and a number of patents protecting the diagnostic test were taken out in France and abroad at the beginning of 1986. Diagnostics Pasteur was awarded the exclusive rights to develop a commercial test kit for HIV-2.

According to the Pasteur Institute, DuPont de Nemours developed a copy of the test kit without authorization.

This is the second time the Pasteur Institute has been involved in a legal wrangle over patent rights to an AIDS virus test kit. A protracted dispute between the Pasteur and the US Department of Health and Human Services (HHS) over who first isolated the HIV-1 virus — and therefore could claim rights to the antibody test kit — was finally resolved in March last year when both parties acknowledged the 'joint invention' of the kit and agreed to share the patent rights (see *Nature* 326, 533; 1987).

Peter Coles

New AIDS institute

Boston

HARVARD University last week announced plans for a university-wide institute to "expand and accelerate AIDS research".

Virologist Myron Essex, currently head of the Department of Cancer Biology at the Harvard School of Public Health, will head the new institute. He will be joined by senior Harvard investigators including William Haseltine, Martin Hirsch and Jerome Groopman.

Essex stressed that an important overall objective of the AIDS institute is to involve more than just experts in virology and immunology. In addition to biological and clinical studies, research centres will be established in epidemiology, policy and education, and international cooperation. Essex emphasized that the enterprise will be more than a new label on existing efforts, saying that he anticipates "think tank sessions", regular meetings among senior investigators and a "high level of interdisciplinary cooperation". At least 70 Harvard faculty members are expected to participate in the institute.

One impetus for the new institute is the desire to help attract more funding from the government, foundations and industry for AIDS research.

Seth Shulman

Commercial venture to speed up AIDS drug tests in California

Berkeley

A NOVEL attempt to accelerate the testing of new AIDS drugs has been launched by the new San Francisco company ViRx Medical Group, founded with backing of \$1.5 million from the Boston venture capital firm of Burr, Egan, Deleage and Company.

ViRx opened its San Francisco clinic last week for patients infected with human immunodeficiency virus (HIV), and now plans to solicit drug-testing contracts from pharmaceutical companies. ViRx boasts of a roster of consultants with expertise in drug development and AIDS virology, including French AIDS researcher Luc Montagnier of the Pasteur Institute.

The company's founders say they are motivated less by profit than by a desire to find effective drugs and make them rapidly available. Bruce Decker, company co-founder and principal consultant, who is himself an HIV patient, says that ViRx hopes to "shave years off the approval process". By building up a pool of several thousand patients whose conditions will have been monitored and recorded, it hopes to streamline the recruitment of subjects for studies. By speeding up efficacy testing, Decker hopes to see more promising drugs gain approval for wide distribution, as AZT did, even before the evaluation process is complete.

The clinic will provide patients with complete blood testing, counselling and monitoring at 30, 60 or 90 day intervals, as

well as advice on the available experimental drugs. Patients participating in studies will be monitored free of charge. ViRx hopes to attract patients by allowing them to choose the drugs they wish to take, rather than conducting placebo-controlled trials which, Decker says, may "no longer be defensible". Instead, ViRx aims to provide matched control groups of patients who have chosen to forego treatment or to take accepted treatments such as AZT.

Decker is not worried that this open attitude toward drug trials could spread the patients into hopeless numbers of categories, at different stages of their disease and with a spectrum of drug histories. He says that promising drugs will bring patients flocking to participate in studies, when it should be possible to persuade them, for the sake of uniformity, to refrain from taking other less promising drugs. Thus Decker says he will have no problem attracting 400 people immediately for a study of dextran sulphate, an antiviral agent currently in vogue.

Companies using ViRx's service may take their test results for approval through the Food and Drug Administration (FDA) or through California's independent approval programme, which has been considering applications since January (see *Nature* 329, 476; 1987). An FDA spokesman says the agency prefers to see placebo-controlled trials, but may approve other controlled trials if they are properly designed.

Marcia Barinaga

Another reminder of Chernobyl

London

TWO years and a day after the reactor explosion at Chernobyl — on 27 April — a leading Soviet atomic scientist died. Official obituaries have not given a cause of death for 51-year-old Academician Valerii Legasov, fuelling speculation that he was another victim of Chernobyl.

Legasov, who was deputy director of the Kurchatov Institute of Atomic Energy, was one of the team involved in investigating the cause of the Chernobyl incident, and in the clean-up that followed. The official death-toll remains at 31, a figure regarded with scepticism by many of the local Kiev residents, according to a recent underground publication. Legasov's death comes at a time when the official media are campaigning to restore public confidence in nuclear power.

In spite of the setback at Chernobyl, the Soviet energy programme remains irrevocably committed to nuclear power. Chernobyl-type RBMK graphite reactors

are to be phased out in favour of the water-cooled VVER model, and plans for new power stations at Minsk, Krasnodar and Chyhyryn have been dropped. Otherwise, though, the programme continues as before. Suggestions that nuclear power stations should in future be sited in remote areas — or, as Academician Sakharov has suggested, deep underground — have been ignored. For the Soviet nuclear programme will eventually be expected to produce not only electrical energy but also hot water for district heating, and the plants will need to be located close to cities.

The campaign to re-establish the good name of the nuclear power industry has made much of the improved training and psychological assessment of power station workers, and the new *Spetsatom* service for phasing out and mothballing obsolete plants. But *Pravda* quotes the local party committee as saying that the managers of the plant are still neglecting vital maintenance work.

Vera Rich

New French government line-up looks promising for research

Paris

THE re-election of François Mitterrand as President of France, followed by the nomination of a new government, has brought several familiar socialists back to key ministries. On paper the changes look good for research and higher education, but the academic community is keeping its champagne on ice. It is still uncertain

several years. He later became director of the Centre National de la Recherche Scientifique. Between 1976 and 1984 he was president of the National Space Research Centre and, from 1981 to 1984, president of the European Space Agency.

Curien's experience of the domestic and European space programmes will, however, have little direct outlet in his new post, as responsibility for space research has been taken away from the research ministry and given to the ministry for posts and telecommunications, now headed by former socialist defence minister Paul Quilès. But Curien is likely to have his hands full. In 1986, just before the socialists lost power, Curien proudly boasted of the creation of 1,400 new research posts and a 14 per cent increase in the state research budget. Much of this effort was undone when, in 1986 and 1987, the Chirac government shifted the emphasis away from direct state spending on research, towards policies to encourage private sector participation.

Ironically, it was Curien who advanced the use of tax incentives to encourage

private sector spending on research and, like his right-wing successors, felt that too few scientists were entering industry. This bipolar attitude to research spending makes it difficult to judge how, with Jospin, he will interpret the president's commitment to science and technology.

At the headquarters of the leading union for academics and science researchers, SNCS-FEN, Curien's appointment has provoked only cautious optimism. "We had Curien as minister for two years", says biochemist Paul Janiaud, the union's secretary, "so we know him. He had some very firm ideas. We hope he will be more open to dialogue than he was the last time." What researchers wish for from Curien is a return to the previous socialist policy of a net annual progression in recruitment of young postdocs — to offset ageing of the scientific community.

Having seen how the French voted during the presidential elections Mitterrand has promised "openness" (toward the political centre) in his government. But it also became clear during the electoral campaign that 'cohabitation' with the right-wing majority was no longer possible. Mitterrand has now chosen to dissolve the assembly, calling for a general election in the hope of gaining a more comfortable majority. **Peter Coles**

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Arnould Baumann.

Hubert Curien, again minister for research.

exactly how Mitterrand will translate his pro-research campaign promises into action (see *Nature* 332, 768; 1988). But few expect an immediate return to the expansive optimism of the socialists' last mandate during the early 1980s.

As widely expected, Mitterrand chose Michel Rocard as Prime Minister. Rocard, whose father, Yves Rocard, was director of the Physics Laboratory at the École Normale Supérieure, was minister of agriculture in the Fabius government from 1983 until he resigned in 1985, having disapproved of proposed electoral reforms. Rocard took only two days to name his government. Instead of the promised 'centrist' cabinet, with an attentive ear tuned to the right-wing majority, the line-up has already been branded 'mitterrandist' and may turn out to be unworkable without a shift in the balance of power.

Mitterrand's faith in education and research as a long-term investment is clearly translated in the new government. High-flyer Lionel Jospin, who has just resigned as head of the socialist party, has been given one of the four ministries of state — education, research and sports. But, in practice, it will be Hubert Curien, who replaces Jacques Valade as minister delegate for research, who will be at the sharp end of government policy.

Curien was minister for research and technology in the Fabius government until March 1986, when the socialists lost power, and so is well-known to the academic community. A physicist specializing in crystallography, Curien lectured at the University of Paris VI for

London Zoo to get £10 million from DoE for new twenty-first century image

London

THE brand of free marketeering zealously advocated by the present British government has caught up with London Zoo — the world's oldest zoological garden. A cash injection of £10 million together with a stiff dose of hard-headed commercial management should rejuvenate the ailing zoo for a financially independent future, says the government. The 162-year-old Zoological Society of London, which runs the zoo and its sister establishment, Whipsnade Park, appears to agree.

Since 1984, when the government stepped in to rescue the zoo from the brink of bankruptcy, the society has depended on an annual grant of £2 million from the Department of the Environment (DoE).

Last week the DoE announced that it is to cease its annual subsidy and will instead give the society a final single payment of £10 million for 'modernization'. As part of the deal, the society's research organization, the Institute of Zoology, will become formally attached to the University of London and will receive an annual grant of £1.3 million through the Universities Funding Council, which under government proposals will next year replace the existing University Grants Committee.

For the institute, which was created in 1977 through the amalgamation of the Wellcome Institute of Comparative Physi-

ology and the Nuffield Institute of Comparative Medicine, this is good news. Income will no longer be largely dependent on such capricious factors as attendance figures at the zoo.

Although the possibility of handing over responsibility for the institute to one of the research councils was considered, it was decided that because the institute has dealings with all four principal research councils, a biased research programme might have resulted if core funding was received from one council only. Additionally, strong informal links already exist between the institute and the university.

For the zoo, the plan is to delegate full management functions to a wholly owned subsidiary company, to be called Zoo Operations Limited and whose chief executive will be Mr Andy Grant, who has previously been involved with San Diego and Philadelphia zoos in the United States. The new company's first job will be to produce an operational plan, which will need to be ratified by the government as part of the conditions for the endowment.

The zoo's administrators say it is too early to go into details of a precise strategy for 'revenue enhancement', but that any changes will be subject to the zoo's charter, and fears that the zoo will become a version of 'Disneyland' are unfounded.

Simon Hadlington

Soviet plans disturb archaeologists

London

AN important archaeological site in the Crimea is being destroyed by bulldozers on behalf of property developers. Local archaeologists (and astronomers) are protesting strongly to the developers: the Academy of Sciences of the USSR.

The site has turned out to be a Byzantine cemetery dating from the fifth to the tenth centuries, and has been under state protection since 1947 when it was declared a conservation area. The main feature of the planned development is a holiday rest home for the academy, which will need a new garage and office block to serve it. The man due to become director of the rest home, Nikolai Zapatskii, claims that the area has also been designated a "test area for scientific equipment".

Archaeologists from the Yalta Historical Museum and the Institute of Archaeology of the Ukrainian Academy of Sciences have attempted a rescue dig but, they say, Zapatskii made the normally hectic schedule of rescue archaeology even more difficult by insisting that they kept to a timetable known only to Zapatskii himself.

This pressure seems to have been a breach of the law. Dr Vadim Mosson, a leading Soviet archaeologist who last week delivered the Gordon Childe Memorial Lecture at London University's Institute of Archaeology, stressed that current Soviet legislation states that no construction work may take place at sites of archaeological importance until a proper dig has been completed at the developers' expense, and until the archaeologist in charge has signed a release.

No action has been taken by the academy to prevent destruction of the site, although individual academicians have visited the site or complained to the central leadership. Rauf Munchaev of the Institute of Archaeology of the Academy of Sciences of the USSR wrote: "Taking into account the unique character of the monument and its great significance for the study of ethnogenesis and cultural history of the peoples of the USSR, no construction work should be possible at this site".

Astronomers have added their complaints to those of local residents and archaeologists. Their concern is that 'seeing conditions' for the new million-ruble telescope planned for the nearby Simeiz observatory will be seriously affected by the light from the new building. G. Khromov of the All-Union Astronomical-Geodesical Society is another of the objectors writing to the academy: "Whatever personal ends the administrative service of the Academy of Sciences of the USSR is pursuing, the interests of our national science should always take first place".

Vera Rich

German university: an education that takes a lifetime

Cologne

A WEST German student finishing his or her last university examination is likely to be three or four years older than a British or French student in the same field. This situation is not new, but it is becoming more worrying in West Germany as 1992 approaches, the year of integration within the European Community.

The Wissenschaftsrat, West Germany's science advisory council, has focused attention on the problem with a study released at the end of April detailing the differences in length of instruction in the humanities and social sciences at West German universities.

The study revealed "astounding differences" in the length of time required to complete a course of study (for a Diplom, the rough equivalent of a US master's degree) in the same field at different universities. In German studies, for example, students in Kiel and Hanover completed their degrees in four semesters fewer than their counterparts in Saarbrücken. A report on the natural sciences will appear at the end of May, and the differences are expected to be just as pronounced in those disciplines.

The Wissenschaftsrat intends that students should use its figures in choosing universities. The resulting shift in attendance, it hopes, will encourage universities to shorten their courses.

Education Minister Jürgen Möllemann

(Free Democrat) has also joined the fray. At the annual meeting of the West German Rectors' Conference here on 10 May, Möllemann said the average study period of seven years would have an "unbearable" effect on West German competitiveness. Möllemann said he would make it a priority to encourage the universities to try to reduce the period from seven years to five.

But Möllemann has a hard fight ahead of him. The unusually long studies in West Germany are due to chronic overcrowding in the universities; to poor planning of examinations by individual faculties; and to the tendency of some faculties to overload their curricula with unnecessary requirements. Kurt Biedenkopf, a former university rector, a member of the Bundestag (parliament) and a national leader of the Christian Democrats, warned Möllemann that if he succeeds in reducing the average length of study to five years, he can expect a "significant increase in unemployment" among university graduates. Staying at university is one way for some students to keep off the dole.

Möllemann also has to contend with the impact of the federal system on West German education. Universities are administered by the Länder (states) in which they are located, which limits the power of the federal government to change the system. **Steven Dickman**

British group makes Soviet deal on seismic station

Moscow

THE British Seismic Verification Research Project (BSVRP) reached a formal agreement with the Soviet Academy of Sciences last Friday to install a seismic station in Central Asia. As under an earlier agreement with the US National Resources Defense Council (NRDC), the objective is to gather data contributing to discussions on monitoring a comprehensive test-ban.

Although the British agreement is on a smaller scale than that with NRDC, and does not require reciprocal access to British testing sites, the agreement is taken as a mark of continuing Soviet willingness to have seismic instruments within the borders of the Soviet Union.

Under the agreement, the British group will be able to establish from July a seismic station at Garm in Tadzhikistan, 1,000 km from the Soviet test-site in East Kazakhstan. The instrument will be a Güralp broad-band seismometer recording the

vertical component of motion and will be built in Britain.

The intention is that the new station should be established and manned by British scientists for the first few months, after which Soviet scientists will help with routine data-gathering and maintenance. The agreement requires that all data collected shall be made available to both sides.

The signatories for the British group, which has been supported so far by a grant of £17,500 from the Rowntree, Cadbury and Puckham trusts, were Aftab Khan (University of Leicester), Frank Barnaby (writer and consultant on arms control) and Dai Davies (a previously editor of *Nature*).

Those responsible on the Soviet side for the negotiations and the agreement were E.P. Velikhov, deputy president of the Soviet Academy of Science, and M. Gokhberg, deputy director of the Soviet Institute of the Physics of the Earth. □

Protests against nuclear power stations growing in Japan

Tokyo

ACTIVISTS against nuclear power armed with radiation-monitoring equipment on Wednesday last week (11 May) chased a convoy of nuclear fuel trucks halfway across Japan. The action was part of a blossoming anti-nuclear power movement that is worrying government and industry.

Until the Chernobyl disaster in 1986, there had been little public protest against nuclear power in Japan. Just months after the accident, the Ministry of International Trade and Industry (MITI) announced plans to build 120 nuclear power plants by 2030 to supply 60 per cent of Japan's electricity (see *Nature* 322, 399; 1986).

But several recent events have changed the mood. First came reports that some imported foodstuffs, such as nuts, macaroni and tea, contain Chernobyl-derived radioactivity. Although the amounts were in most cases only a few tens of becquerels per kilogram, far below safety limits set in the West and Japan, several food co-operatives banned sale of the products.

Then, in February, thousands turned out to protest against a power modulation experiment at the Ikata nuclear plant in the island of Shikoku. And at the end of last month, nearly 10,000 people staged rallies, lectures and protest marches in Tokyo to mark the second anniversary of Chernobyl.

New nuclear plants open in Japan almost every year, and the percentage of electricity produced by nuclear power is rapidly rising, at present accounting for about a third of Japan's output. As a result, it will soon be necessary to modulate nuclear power plant output to meet hourly fluctuations of demand. At present, nuclear plants run at full capacity and fluctuations are accommodated by conventional power plants. But protesters say the manual operations required to modulate power will increase the risk of a nuclear accident.

Another concern is that huge amounts of nuclear fuel, spent fuel and waste are being shipped around Japan in trucks, often unannounced and through densely populated areas, including Tokyo — leading to last week's scheme to trail a convoy of trucks carrying fuel rods for 900 km from the Japan Nuclear Fuel factory in Yokosuka, south of Tokyo, to a power plant at Matsue city on the coast of the Japan Sea.

But government officials and leaders of the nuclear power industry have risen to the protesters' challenge. At the opening of the Japan Atomic Industrial Forum's annual conference last month, Soichiro Ito, head of the Science and Technology Agency, and Jiro Enjoji, chairman of the

forum, called on the nuclear industry to mount public information campaigns to deter further opposition.

Tokyo Electric Power Co., one of Japan's most profitable companies with an annual income of about 500,000 million yen (\$4,000 million), has announced a 20 per cent increase in its public relations budget. And in an unprecedented move last week (12 May), Hajime Tamura, head

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

"What is going to happen mummy? (nani ga okoru no okaasan) asks this poster, designed by an anti-nuclear group in Akita Prefecture headed by Giro Hirayama. (Kyodo Photo.)"

of MITI, summoned the presidents of Japan's nine electric power companies to urge them to do more in the field.

Tamura is concerned that the protest movement has spread from the vicinity of power plants to Japan's major cities. He says the protests are not based on expert knowledge but on simplistic slogans such as "nuclear power is dangerous" and "there is a surplus of electricity".

But the campaigns are becoming increasingly sophisticated. Activists have designed an appealing poster of a mother dressing a child in a protective radiation suit while the child asks "what is going to happen mummy?". The protesters who followed the convoy of trucks from Yokosuka to Matsue carried radiation-monitoring equipment and claim that, when the convoy pulled into a service station on Wednesday night, nearby radiation levels were more than 100 times that of natural background.

With the government planning to begin the airlifting of reprocessed plutonium from Europe to Japan in the 1990s, it seems certain that opposition to nuclear power will continue to grow.

David Swinbanks

US/Soviet accord

Washington

A YEAR-LONG Soviet-American conference on the global causes and symptoms of the greenhouse effect, to be conducted by electronic mail, was inaugurated last week with an exchange of computer messages between Moscow and Washington. Roald Sagdeev, director of the Soviet Institute for Space Research, and ex-astronaut Russell Schweickart were in Washington to congratulate, over a telephone line, Walter Roberts, co-chairman of the US side, who had gone to Moscow to receive their greetings.

Fourteen American and ten Russian participants will be able to exchange their ideas on greenhouse warming through a specially created electronic mailbox on a commercial network. The European Space Agency provided the electronic link for the Soviet side, and has three participants. On the question of possible censorship Sagdeev said both sides would send whatever messages they thought fit until their computer screens were interrupted with a "warning from Big Brother".

David Lindley

Global databases

London

THE International Geographical Union last week agreed to take the first steps towards establishing a coordinated international strategy on the gathering, handling and dissemination of information on global processes. More than 50 experts from a dozen countries, including the United States, the Soviet Union and China, met in Hampshire, England, to discuss how the international scientific community can best use the huge quantities of information being generated through techniques such as remote sensing by satellite. Computerized databases have already been established by several international agencies, but more thought is needed about how best to go about setting up and exploiting databases if the international community is to derive the maximum benefit from them. The union concedes that the problems associated with setting up internationally accessible global data bases are immense, but have to be tackled.

Following the five-day meeting last week it was agreed to establish a working group to consider the legal issues of control, ownership, liability and copyright regarding the construction, maintenance and use of global databases. Efforts will also be made to develop standards for interchange of data bases between scientists and international agencies in the areas of data quality and formats. The union will start to identify what moves will be needed to improve accessibility to global data and its usefulness to people working in programmes related to the global environment.

Simon Hadlington

Spill raises questions about radioactive materials in US

Boston

THE chance discovery of a small spill of radioactive material in a laboratory near Boston has raised questions about regulation of the use and disposal of low-level radioactive materials in the United States.

The spill was discovered when a radioactivity alarm was triggered by a truckload of trash about to be incinerated at a Boston dump. A company receipt in the contaminated garbage led state Department of Health investigators to Marshall Diagnostics, a small medical diagnostics company. There they found that six staff had unknowingly been working in the middle of a minor radioactivity spill, and had carried contaminants on their clothes into their homes and cars.

Louis W. Mead, the company's co-owner, had spilled approximately two ounces of tungsten 178 — a radioactive material with a half-life of 21 days — in his laboratory. Mead, who explained he had been "too tired to think clearly", decided not to report the spill, and tried instead to clean it up himself with bleach and wet towels.

Cleaning materials, all unmarked, were then mixed in with the rest of the laboratory's garbage. Now, with the spill discovered, Mead's laboratory has been

sealed off until the autumn when the radioactivity will have decayed to harmless levels. The laboratory shutdown has forced the company out of business.

The chance way the spill was discovered has raised questions about the efficiency with which the use of low-level radioactive materials is monitored. State officials admit they have too few staff to regulate effectively the 900 or so companies licensed by the state and are not even sure how many licences are still valid.

Massachusetts is not typical in choosing to regulate low-level radioactive materials itself; most states leave arrangements to the Nuclear Regulatory Commission (NRC). Karl Abraham, NRC spokesman, stressed that all of the 13,000 institutions licensed by NRC and the additional 10,000 institutions licensed by states to handle low-level radioactive materials, including hospitals, universities and private laboratories, are required to dispose of these materials in appropriately marked containers and ship them to appropriately licensed burial grounds.

But with only 260 NRC inspectors nationwide overseeing the handling of all nuclear waste, some of them located permanently at nuclear power plants, compliance can hardly be guaranteed.

Abraham acknowledged that many of

the institutions using small quantities of low-level materials are inspected only every four or five years. Even at larger institutions, however, compliance does not always meet expectations. A little over a year ago, Harvard University was cited for a dozen violations in its handling of low-level radioactive materials in a surprise NRC inspection and was fined \$10,000. Inspectors found that researchers had disposed of low-level radioactive materials down the drain.

Seth Shulman

Zero resistance in Japanese research

Tokyo

JAPANESE researchers have developed a technique for making the highest-temperature phase of the new bismuth superconductors which yields stable samples of quite high purity. The bismuth-strontium-calcium-copper oxide superconductors discovered by Hiroshi Maeda of Japan's National Research Institute of Metals a few months ago are now believed to have three phases with zero-resistance temperatures (T_c) of about 80 K, 90 K and 105 K (see *Nature* 333, 153; 1988). But it has been impossible to prepare pure samples of the highest-temperature phase, and the T_c of mixed-phase samples containing the high- T_c phase tends to decrease when an electric current is passed through the samples.

Now a joint group of researchers at Kyoto and Okayama universities claim to have prepared 85–90 per cent pure samples of the 105-K phase by incorporating small amounts of lead in the samples and by using metal oxalates rather than oxides in the preparation process. The samples show zero resistance at 107 K and maintain this transition temperature even when currents of 2–200 mA per cm² are repeatedly applied.

According to Monday's issue of *Nikkei Superconductor*, Dr Mikio Takano of Kyoto University's Institute for Chemical Research believes that the lead partially replaces bismuth in the samples, but this has yet to be proved. Most of the lead is lost during preparatory heating (inductively coupled plasma atomic emission spectrophotometry indicates a Bi:Pb:Sr:Ca:Cu ratio of 65:5:100:85:180), but if the lead is reduced the percentage of the high-temperature phase decreases. On going to press, details were to be reported on Wednesday at a conference in Tokyo.

Almost simultaneously with this announcement, NEC Corporation claimed to have developed a thin film of the high-temperature phase which also shows zero resistance at 107 K — the highest temperature so far reported for a thin film of bismuth superconductors. But in NEC's case, the purity of the phase in the film is only about 50 per cent.

David Swinbanks

Advances in superconducting fibres using lasers

Berkeley

STANFORD's Center for Materials Research last week announced an advance that may lead to superconducting wires which are flexible and capable of high current density at relatively high temperature.

Wires would be made from fibres produced by using a laser to melt the top of a ceramic rod of bismuth-strontium-calcium-copper oxide. A seed crystal is introduced into the melted spot and then drawn out to form a thin fibre containing aligned crystal grains.

Robert Feigelson, director of the Crystal Science Department, said there is no theoretical limit to the length of fibres and that preliminary tests have shown them to be "quite flexible". The technique has produced single crystals 8 mm in length. The dense, aligned structure and longer length of the crystals promise to minimize the weak links at grain boundaries that limit current flow in ceramics with non-aligned crystals.

A 1-cm-long fibre produced by the laser technique has sustained a current flow of 30,000 amps per cm² at 4 K. Feigelson

says, and probably would have done even better if a contact had not failed. While that current density is greater than has been reported in bulk materials at 4 K, the important question of current density at 77 K has not yet been addressed.

To be commercially useful, high-temperature superconductors must support current densities close to 1 million amps per cm², a goal that has been achieved in thin films, but has not yet been approached in fibres. Researchers at AT&T Bell Laboratories recently reached a critical current density of 17,000 amps per cm² with an yttrium-barium-copper oxide fibre in which grains were aligned by melt-textured-growth.

But Sungho Jin at Bell Laboratories called Stanford's laser method an important advance, and agreed with the Stanford group that improvements in grain alignment in fibres should lead to commercially useful critical current densities.

Meanwhile, Stanford attorneys are seeking a patent for the laser process, which will be published this summer in *Science*.

Marcia Barinaga

West German clinical research seeks new directions

Munich/Bonn

REBUILDING medical research after the Second World War has been much more difficult than making the West German economy boom. Especially in clinical fields, West German medicine has not yet recaptured its reputation at the turn of the century. But the authorities are still trying.

The deficiency in clinical science has been a favourite theme of the Wissenschaftsrat, West Germany's science advisory council, for several years, but the federal government finally took heed in February, creating a programme to establish "clinical research groups" at university hospitals around the country.

The research ministry (BMFT) will spend as much as DM30 million per year in a five-year programme to establish up to 40 groups in a variety of clinical disciplines. Groups will include both physicians and basic researchers, to be paid by the federal government for six to nine years, when the appropriate regional government (*Land*) will take over the costs.

But the initiative will not solve long-term problems, nor pull good researchers out of the air. Although the budget will add significantly to the DM50–DM70 million granted in support for clinical research by the Deutsche Forschungsgemeinschaft (DFG), neither sum is large.

The lack of a broader programme can be traced to the structure and history of the university hospital system. Eberhard Buchborn, director of Munich University's city-centre clinic, says that clinical research is not organized into clinical research units as in Britain and the United States. The director of a clinic is responsible for patient care as well as research, and can divide up resources as he or she sees fit. But many physicians are not available for research because of teaching duties. Buchborn said the student/faculty ratio can be as high as 25:1, compared with 1:1 in the United States.

As clinic heads are also entitled to share in the clinic's profits from private patients, there seems also to be little incentive to spend them on basic research. As one biologist put it, "Why should a director put the profits into research when he can put it into his villa?" The salary of university clinic directors is between DM100,000 and DM150,000, but may be augmented by up to DM1 million a year, especially at clinics providing lucrative services such as radiology or surgery.

Buchborn nevertheless disputes the charge that clinic directors deliberately skimp on research spending. He says that, to become a clinic director, a physician must have done many years of research,

which is the primary concern of the faculty that chooses the director. One check in the system is that a professor may not become head of the clinic awarding a postdoctoral qualification (*Habilitation*).

Among other reasons for the state of clinical research is that noted in a 1986 study by the Wissenschaftsrat, which found that 40 per cent of medical professors were forced to leave their posts in 1933, when the Nazis came to power. The study found that efforts to rebuild the West German health system were understandably concentrated on patient care, to the neglect of research. Buchborn adds that the university reforms of the 1970s, which led to the splitting of many previous

clinics, made it difficult to create a critical mass of researchers.

Even so, there is hope that good will emerge from the government initiative. Several West German research groups have international reputations (Hannover in liver transplantation, ultrasound to destroy kidney and gall-bladder stones in Munich, for example), showing that obstacles can be overcome.

Unfortunately, the results of the programme are not expected to be uniform across West Germany. Even the three "model" groups, in Münster, Würzburg and Giessen, on which the programme was based, did not get approved without a fight after the trial period ended. The DFG, which is administering the programme, expects a struggle, especially in the northern Länder, in winning support after federal funding runs out.

Steven Dickman

India becoming more attractive to companies and graduates

New Delhi

THE 'brain drain' of qualified scientists from India is in decline, according to Ministry of Education estimates. The number of science and engineering graduates leaving the country annually is now less than 1,000, or 2 per cent of the annual output. Only 10 per cent of graduates from the Indian Institutes of

uates who might otherwise have migrated.

For one group of graduates, however, the government has decided to take more stringent action. Too many medical graduates have been leaving the country, so doctors "belonging to scarce categories" will not be allowed to emigrate, and medical graduates will be allowed to go abroad for training only if no suitable facilities are available. The Health Ministry is now helping new medical graduates to set up clinics in villages by arranging low-interest bank loans.

Fully owned subsidiaries of overseas high-technology companies are also helping to provide opportunities for Indian graduates staying at home. Comparatively low salary rates, together with the availability of qualified graduates, are attracting pharmaceutical and computer companies in particular.

Halting the loss of home-grown talent is one thing, but can the brain drain be reversed? The government now has several schemes under which Indian technologists get preferential treatment if they return to set up new enterprises in India. So far, the TOKTEN programme (for Transfer of Knowhow Through Expatriate Nationals) has successfully brought 170 expatriates to India for brief stays to work on specific projects.

The manpower division of the Council of Scientific and Industrial Research maintains a register of Indians wishing to return home. In the past 30 years, 24,000 Indians educated abroad have found jobs through this register. New registrants are being added to the list at a rate of about 40 per year, and the government hopes that the number will increase.

K. S. Jayaraman

The bad news is, they're all Sikh militants...



Technology migrated last year, compared with 20 per cent four years ago and nearly 40 per cent in 1977.

Factors behind this partial plugging of the brain drain are said to include increased opportunities for self-employment, improved working conditions and higher salaries. And the Department of Science and Technology says that the newly created agencies in high-technology areas such as microelectronics, computers and biotechnology have created job opportunities for fresh grad-

Progress towards gene therapy

SIR—I would like to draw attention to two points about the progress towards gene therapy that were not dealt with adequately in D. J. Weatherall's comments (*Nature* **331**, 13; 1988) and your leading article (**331**, 100; 1988).

I believe that before somatic cell gene therapy is used in human subjects, it will be necessary to evaluate its safety in hundreds of laboratory animals, each of which is allowed to live until death from natural causes. This is because carcinogenesis is likely to be the principal risk of DNA insertion into somatic cells. Some risk of modification of the expression of an oncogene, using that term in its broadest sense, is undoubtedly a potential risk of this procedure. (Disruption of other types of genes in the somatic cell is not likely to have serious effects on the whole organism.) As the development of cancers is age-dependent, animals subjected to these treatments must be allowed to grow old, and as the risk is likely to be only a few percent, large numbers of animals will need to be evaluated.

Somatic cell gene therapy has much in common with tumour chemotherapy. In both situations, treatment is initially used only on patients with serious and life-threatening disease who have much to gain and little to lose. With both treatments, the principal risk is carcinogenesis.

Most of the debate about germ-line gene therapy has focused upon ethical issues. I happen to agree that it would be unwise for the human race to accept germ-line gene therapy for any reason. But there is a much more compelling reason for dismissing germ-line therapy of human genetic diseases: the procedure is quite simply unnecessary. Insertion of DNA into fertilized eggs (or for that matter, into germ cells) would need to be performed *in vitro*. Almost all couples at risk of producing children with serious genetic diseases face a 1 in 2 (autosomal dominant) or 1 in 4 (autosomal recessive or X-linked recessive) risk. The *in vitro* procedures would always involve fertilizing multiple eggs or evaluating multiple germ cells. Among these would be some that would produce affected offspring and other that would produce normal offspring. Tests to distinguish these two classes of fertilized eggs or germ cells would be needed before gene insertion could be contemplated. (Insertion of additional DNA into an already normal fertilized egg or germ cell would be quite untenable.) Gene therapy would become unnecessary because a normal fertilized egg could be chosen for implantation or a normal gamete used for fertilization.

I can see a possible future role for diagnosis of genetic diseases in eggs fertilized *in vitro* or in germ cells to be used for *in vitro* fertilization, but I cannot

see any role for gene therapy of defective fertilized eggs or germ cells.

The subject of gene therapy is likely to be debated publicly for some years. It is important for scientists to realize that they must spend time explaining their plans to the general public and for members of the public (and the politicians) to realize that they must invest time and effort in understanding what the scientists are talking about if society is to make sensible decisions.

DAVID M. DANKS

*The Murdoch Institute,
Royal Children's Hospital,
Parkville, Victoria 3052, Australia*

Arms go south

SIR—The view of Nobel laureates that the 'conventional' wars have continued to disrupt the Southern Hemisphere while the "Northern Hemisphere has known 40 years of peace since the bombing of Hiroshima and Nagasaki" is too simplistic (*Nature* **331**, 382; 1988). Is it not the major powers in the Northern Hemisphere that supply the conventional weapons to the feuding southern nations? If the southern nations deserve condemnation, why have the weapons factories in the United States, the Soviet Union, Britain, France and Italy been functioning at full capacity for the past 40 years or so?

It is also amusing that the Nobel laureates did not criticize France which exports 75 per cent of its massive \$11 million aerospace production.

SACHI. SRI KANTHA

*Laboratory of Marine Biochemistry,
University of Tokyo,
Bunkyo-ku, Tokyo 113, Japan*

Funding of science

SIR—A reason why the differences between Mayr and Weinberg on reductionism and on priorities in the funding of the sciences (*Nature* **330**, 433; 1987 and **331**, 475; 1988) cannot be resolved by clearing up the semantics (Mayr) or by an attempt to "express myself a little more clearly" (Weinberg) is because they are talking about two different enterprises: Mayr, with customary clarity, is talking about hierarchical or vertical reductionism (that an increased understanding at a fundamental level will not dispense with the necessity to understand the emergent properties of greater complexities in different ways), whereas Weinberg, while agreeing with that, is writing (mainly) in his response about horizontal or chronological theory-displacement—admittedly of an elementary kind, rarely encountered, that what is true in T1 is straightforwardly subsumed in T2, in some kind of

smooth transition. With that understanding of theory displacement, Weinberg is able to exploit the ambiguity of the word 'fundamental' to mean *both* 'deep' and 'elementary' *and also* 'fundamentally more important than other things'. The ambiguity can be seen in his lecture, in the paragraph beginning, "The case for spending large sums of money..." (*Nature* **330**, 434; 1987).

All this could be happily left to the philosophers of science to sort out in due course, except that large sums of money are indeed at stake (for both the Superconducting Super Collider (SSC) and CERN). Whether such sums should be allocated is a social decision (as well as a scientific judgement of priority) whenever the redistribution of resources through taxation is involved. Consequently, although what Weinberg learnt about the art of congressional testimony was to keep his mouth shut, it is essential for him, and for all of us, to say a great deal more about the competition for resources than that "we don't really know with what the SSC will compete for funds". Congressional committees are one kind of forum where social considerations can be included; maybe *Nature* is another. In any case, the issue will not be decided by setting eliminative reductionism against theory-displacement when both protagonists agree that the former is false and the latter inevitable. What *would* help the debate and the decision would be to make explicit the criteria which they (and the rest of us) actually do adopt when choices of this kind have to be made. Perhaps Weinberg will open his mouth again?

JOHN W. BOWKER

*Trinity College,
Cambridge CB2 1TQ, UK*

Fertility symbols?

SIR—I wish to comment on the News and Views item "Indian maize in the twelfth century BC" by C. Johannessen (*Nature* **332**, 587; 1988) — the date of the title should, of course, have been AD.

I do not understand the Western obsession that everything in Hindu temples and Hindu religious rites is a fertility symbol: rice tillers, sugar canes, coconuts and maize, all of which can be seen in the hands of various deities, were in fact simply intended to indicate the blessing for successful crops. Describing these features as anything else is just as ridiculous as describing 'Mary with baby Jesus in her arms' or sculptures of various Christian saints holding flowers as fertility symbols. Hindus never symbolized sex and fertility, but presented them as frankly as possible.

G.R.K. SASTRY

*Department of Genetics,
University of Leeds,
Leeds LS2 9JT, UK*

What happened to nuclear winter?

Calculations of nuclear winter are become mainstream academic work, and thus less contentious than a few years ago. They may, in this new role, point usefully to solutions of other problems.

THE paper by Schneider and Thompson (this issue, page 221) is not the end of nuclear winter but a proof (if one were needed) that nuclear winter has come of age, and has become an accepted ingredient of academic study. In what other circumstances do people write review articles?

By now, the essence of the problem is widely understood. If there were a nuclear war, many parts of the surface of the Earth would be set alight, the fires would carry large amounts of smoke into the atmosphere, the passage of solar radiation to the surface of the Earth would be impeded, we should all, for a time, feel cold and, seriously, there would be a chance that the climatic consequences would persist for weeks or even months, not just days — long enough to interrupt processes such as photosynthesis on which continued survival depends. Nobody can quarrel with that statement; the practical question is that of what, for a given prospective nuclear war, is the chance that the survivors will have only a short time in which to envy the dead?

The Schneider and Thompson paper derives from one delivered at a symposium in Bangkok arranged at the end of 1986 by the ENUWAR executive of the Scientific Committee on Problems of the Environment (SCOPE), itself an offshoot of the International Council of Scientific Unions (ICSU). It is important that, by then, the threat of nuclear winter had widely seized the public mind for only just over two years, since the publication of the paper by Turco, Toon, Ackerman, Pollack and Sagan in *Science* (222, 1283; 1983) describing the predictions of a one-dimensional atmospheric model of the climatic consequences. That, in turn, had been stimulated by an earlier calculation by Crutzen and Birks (*Ambio* 11, 115; 1982) that nuclear wars would send huge amounts of smoke into the atmosphere, with potentially catastrophic consequences. Soon afterwards, Thompson and Schneider were in print (Thompson, Schneider and Covey, *Nature*, 310, 625; 1984) urging the benefits, in such calculations, of more sophisticated climatic models. The list of references on page 227 will show that others have not been idle.

So what has been learned about and from nuclear winter? The first thing to say is that Crutzen and Birks stood impeccably on solid ground by publishing their

calculation of what might happen if there were a lot of smoke in the atmosphere. But, a year later, a bandwagon had been set rolling, press conferences had been arranged, the paper by Turco, Toon, Ackerman, Pollack and Sagan had been renamed "TTAPS" (presumably in emulation of the respect accorded to the paper on nucleosynthesis by Burbidge, Burbidge, Fowler and Hoyle) and the late Mrs Gandhi had taken to complaining that a nuclear war between the superpowers might destroy others than themselves. The present position, mentioned in passing by Schneider and Thompson, is that ENUWAR has sought to standardize the models people play with by presupposing certain quantities of carbon discharged as smoke into the atmosphere (the basic unit is the Tg, one million tonnes).

What researchers of these questions have not properly understood is the effect of their ratiocinations on the world at large. (Crutzen and Birks were seemingly archless in this respect.) But by the following year, the TTAPS team was planning to make what is called an "impact", and did so. *Nature*, having decided (wrongly) that a different press conference in Washington would make better copy, was probably over-sour in its comments on the affair. But the plain truth is that, three years ago, most ordinary people were scared stiff by the threat of nuclear winter: one might survive the blast, but after that would be only the cold and the starvation.

All this on the strength of a one-dimensional model? The complaint embodied in that question must be understood for what it is. Many one-dimensional models are remarkably predictive. There is no reason why a simple model of the atmosphere should not faithfully reflect its behaviour under the influence of external forcing circumstances. Technically, the chief weakness of the TTAPS model was more probably its understandable omission of consideration of what would happen in the first few days of the aftermath of a nuclear war; even now, Schneider and Thompson have little to say on that question, but that is understandable. There are also technical difficulties about the difference between the formation and effects of clouds rather than of "average cloudiness", still much more easily incorporated into climatic models.

What this implies is chiefly that four years of public anxiety about nuclear win-

ter have not led anywhere in particular. At the outset, people in general were convinced that any substantial nuclear war would be a catastrophe and did their best to arrange that there should be agreements between potential combatants to minimize the chances. The hype of nuclear winter might in principle have extended the circle of those eager for the same outcome, but chiefly served to make people cynical; does it matter if you would be dead even if you survived?

So, inevitably, the issue has become academic in the best sense; professional people think it worth study. Schneider and Thompson, and the host of documents they cite or are unable to cite for lack of space, testify to the interest of the problems nuclear winter has thrown up. What really happens at the edges of clouds? (In this case, would induced precipitation wash out the clouds?) What coupling would there be between the atmosphere and the oceans? Would it always be possible to rely on Le Chatelier's principle for a mitigation of the consequences of simple predictions?

For the time being, at least, the issue of nuclear winter has also become, in a sense, irrelevant. At the back of the mind, most people know that the superpowers are about to tell each other of their impending ratification of the agreement not to aim nuclear missiles at each other from bases in Europe, which will go some way to reduce the immediate risk of nuclear wars such as TTAPS supposes, but there are more urgent reasons why people should be worrying about the refinement of climatic models.

The problem of the ozone layer caused by chemically inert aerosol propellants and refrigerants has qualified as daily-newspaper fodder, but the greenhouse problem will eventually be a harder nut to crack politically, if only because burning of fossil fuel (and even of biomass) is so crucial to economic activity. Even when meteorologists are able to put their hands on their hearts and say that they have detected signs of global warming, it will fall to the model-builders to say what the consequences will be. When that time comes, it is to be hoped that they will respond more confidently than to the issue of nuclear winter with opinions properly hedged with specific qualifications about the ways in which the models are incorrect.

John Maddox

Ecology

Diversity and stability

S. J. McNaughton

THERE is something almost mythological about the hypothesis that greater species diversity in ecological communities is associated with greater community stability. Like Hydra, cutting off one of its heads leads to two sprouting back; like the Gorgons, looking directly at it turns researchers to stone. And, like Tantalus, ecologists who study diversity often seem to have resolution within their grasp, only to see it recede with the next data or theory paper. On page 261 of this issue¹, Moore and Hunt quantitatively address a major theoretical aspect of the problem, first considered in *Nature*^{2,3}, that compartmentation and reduced interaction strength associated with greater diversity can stabilize food webs. Moore and Hunt provide direct evidence for compartmentations and for the reticulate nature of species interactions in ecosystems using data on nitrogen flow through a below-ground food web in an American grassland. Their correlative analysis of 40 food webs⁴ indicates that both connectance and the strength of interactions decline with increasing species diversity, a result that confirms the theoretical prediction⁵ that population stability can exist in complex food webs.

Elton first suggested⁵ that a greater variety of species in a community leads to reduced fluctuations of population densities. But theoretical and empirical studies⁶ since then provide contradictory evidence for this assertion. Rather than looking for simple connections between the two, those studying diversity–stability are now trying to identify conditions under which these two parameters are related, and those under which they are not. This should elucidate the relationship between stability and diversity, which is important as the topic is a synthetic confluence of population, community and ecosystem ecology⁶.

Elton⁵ argued that "...the balance of ... simple communities ... is more easily upset than that of richer ones; that is, more subject to destructive oscillations in populations ... and more vulnerable to invasions." He had six arguments to support this contention: (1) simple mathematical models of population dynamics are inherently unstable; (2) laboratory experiments require habitat complexity for prey and predator to coexist; (3) small islands are more vulnerable to invasion than continental areas; (4) invasions and species outbreaks are more common in the simplified cultivated communities of agriculture than in more complex natural communities; (5) high-

diversity tropical communities are less subject to population-density fluctuations than less diverse temperate and subarctic communities; and (6) the use of chemicals in orchards to control pests greatly simplifies the community and often leads to subsequent population outbreaks due to the elimination of predators.

Robert MacArthur, like Elton one of the most influential ecologists of his generation, introduced⁷ an index of community diversity as a measure of stability. Although couched in population terms, MacArthur's argument that a greater variety of channels in food webs stabilizes flow through those webs has found wider application in community and ecosystem ecology⁸ than in population ecology.

It is, I think, fair to say that the notion that greater community diversity is associated with increased stability was among the most influential beliefs in ecology from the 1960s until the mid-1970s, reaching the status, in some cases, of a "core principle"⁹. Hydra's heads then began to fall under the onslaught of a theoretician, Robert May, who introduced³ a concept from cybernetic theory² and built a powerful theoretical argument⁹ "... that, as a mathematical generality, increased complexity makes for diminished community stability. Even in the natural world — which is no general system — it is not true that population stability is uniformly associated with trophic complexity and faunal and floral diversity. Elton's six arguments ... do not stand up well to close examination ...". Nevertheless, according to May⁹, "Natural ecosystems ... are the product of a long history of coevolution It is at least plausible that such intricate evolutionary processes have, in effect, sought out those relatively tiny and mathematically atypical regions of parameter space which endow the system with long-term stability." These views have dominated the diversity–stability agenda for the past decade.

Stuart Pimm^{6,10} then pointed out that stability, complexity and diversity have come to have many different meanings in ecology. The two principal aspects of stability are the constancy of population densities referred to by Elton and the ecosystem process constancy derived from MacArthur. A key caveat of May's critique³ of (then) conventional diversity–stability ideas was that "... communities, for given average interaction strength and web connectance, will do better if the interactions tend to be arranged in

blocks....".

It is the relations among blocking (the food-web channels of MacArthur), connectance and interaction strength that Moore and Hunt examine in this issue¹. These authors have an ecosystem vantage point, but apply it to the population-stability criteria of May. Their analysis of nitrogen flux through below-ground food webs shows that the bottom of food-web channels, where intake is based on substrate properties, are organized in compartments, whereas predators at the top of the webs link channels because their feeding habits are based on prey size and morphology. Moore and Hunt also find that there is temporal compartmentation because of the distinct seasonal abundance patterns of substrates available to the organisms at the base of the food web. Moreover, there is habitat compartmentation within the under-ground food web because of the aquatic nature of the bacteria-based channel, whereas other channels are less tied to free soil water. These data suggest that communities are highly blocked, as May concluded they must be to be stable, except at the top of the food web.

Previous analyses of food webs had shown that unblocked webs have properties suggesting greater fragility as diversity increases¹¹. Applying their food-web source and habitat compartmentation criteria to food webs compiled by Briand⁴, Moore and Hunt¹ find that the number of channels increases with diversity, so that compartmentation increases as the community becomes more complex. Moreover, food-web connectance and average interaction strength declines with the number of energy channels. These data suggest that real food webs do have many of the properties of the rare parameter space that May thought might be characteristic of natural communities.

Ecologists divided into theoretical and empirical camps sometimes seem to speak different languages, but the progress of diversity–stability hypotheses over the exactly 30 years since Elton's conjectures indicates that those camps can interact in a highly productive fashion. What is now needed, I believe, is a rigorously experimental approach to the problem. A fruitful experimental system might be successional microbial communities like those studied some time ago by

1. Moore, J.C. & Hunt, H.W. *Nature* **333**, 261–263 (1988).
2. Gardner, M.R. & Ashby, W.R. *Nature* **228**, 784 (1970).
3. May, R.M. *Nature* **238**, 413–414 (1972).
4. Briand, F. *Ecology* **64**, 253–263 (1983).
5. Elton, C.S. *The Ecology of Invasions by Animals and Plants* (Chapman and Hall, London, 1958).
6. Pimm, S.L. *Nature* **307**, 321–326 (1984).
7. MacArthur, R. *Ecology* **36**, 533–536 (1955).
8. Margalef, R. *Perspectives in Ecological Theory* (University of Chicago Press, 1968).
9. May, R.M. *Stability and Complexity in Model Ecosystems* (Princeton University Press, 1973).
10. Pimm, S.L. *Food Webs* (Chapman and Hall, London, 1982).
11. Yodanis, P. *Nature* **284**, 544–545 (1980).
12. Woodruff, L.L. *J. exp. Zool.* **12**, 205–264 (1912).

Woodruff¹². Applying modern analytical techniques and theoretical concepts to his hay infusions could provide an excellent, synthetically combined approach to many community problems that are intractable at other scales. I have tried unsuccessfully to interest graduate students in this approach, but Woodruff's 75-year-old work seems to hold little current charm. Nevertheless, his system seems to lend itself to the deconstruction and reconstruction necessary for an experimental attack on the functional consequences of community organization. Moore and

Hunt, like all but a few ecologists⁶, infer stability properties from community structure. The internal subdivision of food webs appears to be a critical property if more diverse webs are to be stable. Combining a tractable experimental system with modern technology and thought might allow a synthesis of correlative inference and experimental evidence. □

S.J. McNaughton is Professor of Botany at Syracuse University, Syracuse, New York 13244-1220, USA.

Radioastronomy

An eclipsing binary pulsar

F. Graham-Smith

MOST of the 400 pulsars that have been discovered have rotational periods of 0.1–4 seconds. Unlike most stars, very few of these rapidly rotating neutron stars have been found in binary systems with an orbiting partner; it is generally supposed that the supernova explosion accompanying the collapse of a massive star to form a pulsar is large enough to disrupt most binary systems. But millisecond pulsars, whose periods are less than about 10 milliseconds, are typically in binary systems with other condensed stars, either white dwarfs or neutron stars, as partners. The discovery by A.S. Fruchter, D.R. Stinebring and J.H. Taylor, reported elsewhere in this issue (*Nature* 333, 237–239; 1988), of a binary millisecond-pulsar system in which the partner is large enough to eclipse the pulsar, cutting off the characteristic radio pulses for a tenth of its orbit, is therefore a great surprise.

In a plot of pulsar period P against the slowdown rate $\dot{P}$, as shown in the figure, the youngest pulsars appear at the top left; typical young pulsars are the Crab and Vela pulsars whose respective ages are about 1,000 and 10,000 years. As they slow down, pulsars move downwards and to the right, as shown by the sloping track. During their lifetime of about 10^7 years, the slow-down process becomes less effective because the pulsar magnetic field decays, so that the track becomes more vertical. Eventually their radio emission stops or becomes too weak to detect; at this point they cross the 'death line' shown in the lower right.

Where then can the millisecond pulsars come from, located as they are in the lower left of the diagram? The clue to this lies in their binary partnerships. A binary pulsar which has crossed the death line can be resurrected; its partner can transfer

angular momentum to it, spinning it faster than in its most youthful days. The process of mass transfer which can speed up the rotation to over 600 revolutions per second is already observed in the X-ray binaries, where material falling onto a neutron star is heated by gravitational energy to 10^7 K or more. The rejuvenated radio pulsar, however, conceals its true age, because its magnetic field has already decayed from about 10^{12} gauss (10^8 tesla)

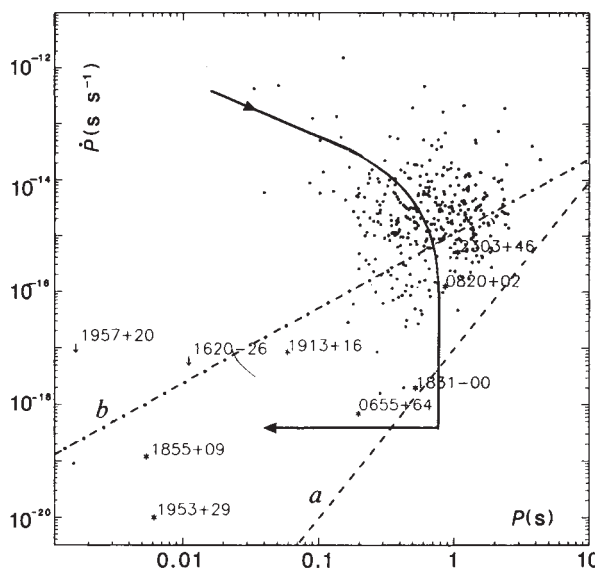
the binary could date back before the neutron star was formed; this implies that a comparatively small amount of mass was expelled in the supernova explosion so that there was no disruption. Most current models lead us to expect that the binary partner should be a white dwarf. This is easily tested, because the orbital period and the changing Doppler shift of radio lines during the orbit can be combined to give information on the masses of the components. The partner of the M4 binary, for example, is clearly a white dwarf with a mass of about 0.3 solar masses.

The outstanding feature of the newly discovered binary system, reported in this issue, is the occultation (eclipsing) of the pulsar by its companion. This immediately shows that the binary orbit is seen approximately edge on, so that the mass calculations are reasonably certain. The companion turns out to have the unusually low mass of only 0.02 solar masses. The occultation, however, lasts for one-tenth of the orbit; it can easily be shown that the companion must be very large, with a diameter about three-quarters that of the Sun, instead of about 1 per cent as expected for a white dwarf.

This problem is solved by a further observation of this extraordinary object. Immediately before and after the occulta-

tion, the radio signals are delayed in a way which is characteristic of passage through ionized gas. Evidently the occultation is caused by an extended hot atmosphere surrounding the tiny white dwarf. There is even an observable difference between the beginning and end of the occultation, showing that the ionized atmosphere is not spherically symmetrical; the appearance is comet-like, with a more extended atmosphere on the trailing side.

Fruchter *et al.* speculate, very reasonably, that they are seeing the evaporation of the white dwarf by the pulsar. This would, incidentally, provide another model for the creation of solitary millisecond pulsars, but attention is naturally focused at this time on the mechanism by which the pulsar can heat the white dwarf. The total energy available from the pulsar rotation is probably about 100 solar luminosities, but it seems unlikely that it is the radiation from the rotating magnetic dipole that is directly responsible. The exciting proposal is that the companion is heated by a stream of high-energy particles, perhaps with the energy of cosmic rays, that are generated in the rapidly rotating pulsar. □



Evolutionary track of pulsars (bold line). Period P and slowdown $\dot{P}$ for 361 pulsars. Catalogue numbers of millisecond and binary pulsars (*) are given. a, Death line; b, spin-up line.

to about 10^9 gauss. The strength of the field, and hence the age of the pulsar, is revealed by the slow-down rate.

There now seem to be several ways in which a binary star system can evolve and produce an observable millisecond pulsar. The recent discovery by A.G. Lyne *et al.* (*Nature* 332, 45–47; 1988) of a binary pulsar in the globular cluster M4 supports the proposal that the binary pair could be the result of a near collision of a neutron star and a red giant. Alternatively,

F. Graham-Smith is Astronomer Royal and is at the Nuffield Radio Astronomy Laboratories, Jodrell Bank, Macclesfield SK11 9DH, UK.

Mathematics

The beat of a fractal drum

Ian Stewart

EVERYTHING in the Universe vibrates. Light from the most distant galaxy and the sound of a nightingale are brought to us by vibrations — of the space-time continuum and the atmosphere, respectively. The tone of a Stradivarius violin, and the stability of a car wheel, depend on how they vibrate. And each object has not just one, but a whole range of characteristic resonant frequencies of vibration. Ancient Chinese bells were designed to have two especially loud natural frequencies: which one sounds depends upon where the bell is struck. Michel Lapidus (University of Georgia, Athens) and Jacqueline Fleckinger-Pellé (Université Paul-Sabatier, Toulouse) now prove in *Comptes Rendus de l'Académie de Science, Paris* (306, 171–175; 1988) an important conjecture about the vibrational frequencies of objects that have fractal boundaries.

Harmonics

The fundamental frequency of a violin is determined by the tension in the string. But it can also generate harmonics of the fundamental frequency — vibrations that occur twice as fast, or three times, or four times. . . . Here the pattern is easily described: the possible frequencies follow the series of whole numbers 1,2,3. . . . But more complicated shapes produce more complicated series of frequencies. Even the analysis of the vibrations of a circular disk, modelling the head of a drum, leads to sophisticated questions about Bessel functions.

Despite these complexities, Hermann Weyl discovered that all vibrating shapes have features in common — at least, if they have smooth edges. In particular, the statistical distribution of the characteristic frequencies obeys regular laws, to be described in more detail in a moment. In 1980 Michael Berry (*Proc. Symp. Pure Math.* 36, 13–29; 1980) conjectured, on physical grounds, that a more accurate version of Weyl's results should be true, and that it should apply not just to the smooth shapes envisaged by Weyl, but to shapes with fractal boundaries. Now this Weyl-Berry conjecture has been proved — in slightly modified form — by Lapidus and Fleckinger-Pellé.

Mathematically, small-amplitude vibrations of any medium are described by the wave equation. This was originally devised in the eighteenth century by Leonhard Euler in a study of musical instruments, but soon Joseph-Louis Lagrange extended it to sound waves, and the floodgates burst. The wave equation became perhaps the most important of all the equa-

tions of mathematical physics. Its mathematical form, $d^2f/dt^2 = \nabla^2 f$ involves the laplacian differential operator ∇^2 , defined by $\nabla^2 f = \partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 + \partial^2 f/\partial z^2$.

The characteristic frequencies of a vibrating shape correspond to the eigenvalues λ of the laplacian; that is, the solutions of the equation $\nabla^2 f + \lambda f = 0$. In this equation you should think of the boundary as being held fixed while the rest of the shape vibrates, just as in a drum or a violin string.

Weyl's formula describes the asymptotic properties of the vibrations at high frequency. Specifically, let $N(\lambda)$ be the number of characteristic frequencies less than a given value λ . How does $N(\lambda)$ behave as λ becomes large? The result is that $N(\lambda)$ is approximately proportional to $\lambda^{k/2}$, where k is the dimension of the vibrating object. The constant of proportionality depends on the volume of the vibrating shape.

Here 'approximately' means that the ratio of the true answer to that given by Weyl's formula tends to 1 as λ tends to infinity. But that gives only very coarse information. To improve Weyl's formula we must ask how large the error can become. Berry's idea is that the error depends on boundary effects, and should be of the same order of magnitude as $\lambda^{d/2}$, where d is the dimension of the boundary. For example, a drum has a circular head, for which $k = 2$ and $d = 1$. So $N(\lambda)$ is

approximately proportional to λ , with an error no worse than $\lambda^{1/2}$.

Weyl's proof of his original formula assumed that the boundary of the shape is mathematically nice: smooth, or at least smooth except at finitely many corners. But Berry's physical reasoning appears to apply equally well when the boundary is irregular. Indeed Berry suggested it should be true when the boundary is fractal, in the sense of Benoit Mandelbrot (*The Fractal Geometry of Nature*, Freeman, San Francisco, 1982).

Fractal snowflake

The archetypal fractal is the snowflake curve. Begin with an equilateral triangle. To each edge add an equilateral triangle one-third the size. Now repeat indefinitely, adding ever-smaller triangles. How does a drum shaped like a snowflake vibrate? According to Berry, much like a drum with a smooth rim, until you look at the fine detail, the high-frequency vibrations that penetrate into tiny crevices of the boundary. Here, more vibrations should be possible, because fractal objects have lots of ever-smaller crevices. So the number $N(\lambda)$ should be bigger. Instead of missing Weyl's formula by about $\lambda^{1/2}$ (as for a circular drum), a snowflake should miss it by more. Berry argues that the error should be about $\lambda^{d/2}$ where d is the dimension of the boundary of the snowflake.

What do we mean by the dimension of the boundary, when the boundary is an irregular fractal? Berry's conjecture is that it should be the fractal dimension, or Hausdorff-Besicovitch dimension. This concept, which is fundamental to the analysis of fractals, measures the

Discovery of a new species of lemur

THESE lemurs were first observed in 1985 by Bernard Meier and Yves Rumpler in the Hanomafana region of Madagascar. Last month it was confirmed that they belong to a new species, *Haplemur aureus*. Since their discovery, these monkeys have been investigated by a team supported by the World Wildlife Fund and Duke University in North Carolina. The team captured two specimens and chromosome analysis confirms that the animals are members of a new species. *H. aureus* weighs about 2 kilograms and is monogamous, each couple living in a territory of 15–20 hectares of forest. The offspring are born in December and live with the parents for 2 years before leaving to begin an independent existence. Adult monkeys live off freshly-grown leaves. Because these lemurs are so rare — only 500 individuals of *H. aureus* are believed to exist — the Madagascan authorities plan to turn this area of forest into a wildlife reserve. Photograph: Jean-Christophe Peyre/Gamma. □

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

behaviour of the snowflake under changes of scale. In particular it need not be a whole number, unlike the more usual concept of dimension. Here d is about 1.2618. So the error in Weyl's formula should be of the order of $\lambda^{0.6309}$, compared with $\lambda^{0.5}$ for a drum with a smooth rim.

This may sound like an exercise in abstract pure mathematics — art for art's sake — but it is not. Many natural objects are better modelled by fractals than by smooth surfaces: the vibrations of water in a lake with an irregular edge, or oscillations of the entire Earth, as studied by seismologists to detect earthquakes and analyse the Earth's interior. The acoustic properties of a concert hall depend on its irregularly shaped walls — beloved of modern architects and having some theoretical justification as improving the acoustic responses.

But Berry's conjecture is false. This was shown by J. Brossard and R. Carmona (*Commun. math. Phys.* **104**, 103–122; 1986). That might seem to be the end of the story, except that the physical intuition behind the conjecture still looks very appealing. Lapidus and Fleckinger-Pellé show that the conjecture can be put right if, instead of the Hausdorff–Besicovitch dimension, the less familiar Minkowski dimension is used.

The Hausdorff–Besicovitch dimension describes how the ' d -dimensional volume' of an object behaves under changes of scale. The Minkowski dimension, in contrast, describes the behaviour of points near the object as well as on it. This has some physical appeal in the vibration problem, because it is the parts of the drum near its rim that vibrate, and not the rim itself. The proof is purely analytical, and extends the classical proofs of Weyl's theorem to the fractal case. It does not yet establish a sharper version of the Weyl–Berry conjecture, that the error can be absorbed into a second 'correction term' in the formula, but it seems likely that this is the case.

The importance of physical intuition for mathematical discovery is clear, and this important extension of Weyl's classic result is an excellent case in point. But it teaches other lessons too. One is the need for proper mathematical rigour, without which it would not have been realized that the most popular definition of fractal dimension is not the one appropriate here. The second is that, although intuition can suggest that certain results are true, it does not always suggest the correct line of attack for a proof — which here required classical analytical methods. And a third is that apparently dead ideas from pure mathematics — here Minkowski's version of dimension, certainly not widely known even among mathematicians — can suddenly spring to life in new applications. □

Ian Stewart is in the Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK.

Protein structure

Helices of antifreeze

Roger H. Pain

ORGANISMS that continue to live when cooled to temperatures at which the cell and body fluids would otherwise freeze, with disruption of biological structures and death, are enabled to do so by their production of so-called antifreeze proteins (AFPs). This depression of the freezing point is caused by inhibition of the formation of large ice crystals rather than by a lowering of the equilibrium melting point. On page 232 of this issue¹, Yang and colleagues describe the crystallization of one AFP, a polypeptide rich in alanine, and determine its structure at an atomic resolution of 2.5 Å, revealing a single helix about 5 nm (50 Å) long. These authors propose a mechanism by which this particular AFP binds to an ice nucleation structure, preferentially along one crystal axis, using unique properties of the α -helix structure.

Many organisms have to survive low temperatures at which their fluids would normally freeze. Certain beetles and pine

former comprises repeating Ala-Ala-Thr units with a disaccharide attached to the threonine, whereas the latter contains no sugar residues but can be high in alanine or cysteine or, alternatively, rich in neither. Significant depression of freezing point has also been achieved with a random copolypeptide containing 65 per cent alanine and 35 per cent aspartic acid.

Freezing takes place by a process of nucleation on non-water particles or on a local water-lattice structure with a regular tetrahedral arrangement of oxygens but an irregular orientation of water molecules (Fig. 1). The hysteresis induced by AFPs has been proposed to occur by their binding to ice nuclei, preventing their growth to large crystals². The problem, then, is to discover the mechanism of binding to nucleation structures that competes effectively with water within a degree of the usual freezing point in such a way as to inhibit growth of the crystal into ice.

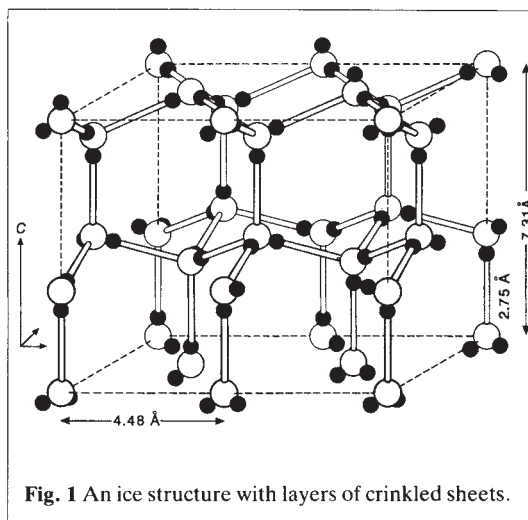


Fig. 1 An ice structure with layers of crinkled sheets.

needles are able to withstand temperatures down to -30°C . In addition to synthesizing polyols that act as true antifreeze, such organisms gain additional protection from AFP macromolecules. Polar fish that can live in sea water in equilibrium with ice at -1.9°C possess proteins in their plasma capable of depressing the freezing point by one degree Celsius but of reducing the equilibrium melting point by only about 0.02°C . For this reason, these molecules have also been called thermal hysteresis factors.

Any proposed mechanism of action for AFPs has to take into account the diversity of their chemical structures. With molecular masses ranging from 3 to 40×10^3 , one group can be classified as glycoproteins and another as polypeptides. The

glycoprotein AFPs contain sugar residues whose hydroxyls could be sterically compatible with the water lattice; reversible inhibition studies with borate ions suggest that this mode of binding does occur. The variation in chemical structure of the different polypeptide AFPs indicates, however, that binding specificity must be sought in a higher level of structure. The AFP from the winter flounder has been shown by circular dichroism to assume a highly helical conformation in solution, the helix content increasing as temperature is reduced³. The structure is in keeping with the strongly helix-forming nature of the component amino acids.

Yang *et al.* in this issue¹ crystallize this AFP and show that it takes up an α -helix conformation over its whole length. Interestingly, they grew the crystals in acetone and stabilized them with isopropanol, which might be expected to stabilize such a helix conformation.

The relevant characteristics of the α -helix conformation are that the polar backbone of the polypeptide is effectively masked and unavailable for external interaction while the side-chain groups are available for bonding, every third or fourth residue being located on the same side of the helix cylinder (Fig. 2). Attention therefore has been focused on those polar side chains such as lysine and glutamate that are appropriately oriented and could hydrogen-bond to water molecules

in a nucleation lattice.

Raymond and DeVries² showed that the binding of an AFP to small ice crystals grown in its presence did not perturb the structure of the ice lattice but resulted in crystal growth being restricted effectively to one dimension, that along the *c* axis perpendicular to the crinkled sheets visible in Fig. 1. Because ice usually grows most rapidly in the direction normal to the

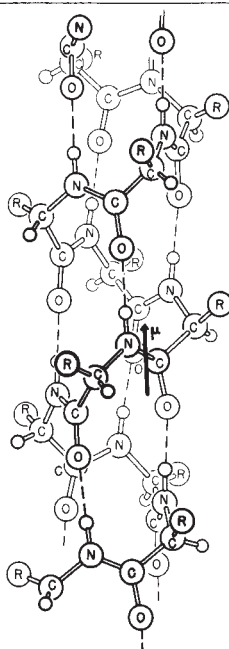


Fig. 2 The side chains (R) of the helix are located on the surface of the cylinder. A dipole is marked for one of the aligned peptide bonds³.

c axis⁴, the AFP is taken to be adsorbed preferentially to the crystal faces parallel to the *c* axis.

Stimulated by their demonstration of the highly regular helical structure of the winter flounder AFP, Yang *et al.* have attempted a synthesis of the above findings by invoking a further property of the α -helix highlighted by Hol *et al.*⁵. Associated with the peptide bond in the polypeptide backbone is a significant electric dipole. When ordered in the α -helix conformation, these dipoles become aligned close to the helix axis (Fig. 2), leading to a resultant dipole along the helix. This dipole is equivalent to an isolated half unit charge at either end of the helix, generating a field that is maximum for between 15 and 20 residues — slightly shorter than the flounder AFP helix.

In ice, the energy difference between different orientations of individual water molecules is presumably very small, as a random distribution persists down to absolute zero. Assuming the activation energies to be sufficiently low, the molecules at the surface of a nucleation lattice could be aligned regularly with only entropic cost. Yang *et al.* propose that the AFP helix, interacting through polar

hydrogen-bonding groups with the ice lattice, induces an antiparallel orientation of water molecules in the field of its dipole. Optimum alignment of water molecules would lead to a resultant lattice dipole inclined at 51° to the *c* axis and at 60° to the *a* axis (see their Fig. 5 on page 236). Maximum interaction of the AFP helix would therefore occur if it were aligned with this dipole.

Were such binding and alignment to occur, it would account for the needles of ice found in the presence of this type of AFP² and for the inhibition of crystal growth leading to ice formation. There is,

however, going to have to be some ingenious and elegant experimentation to substantiate (*pace* Popper) this attractive hypothesis. □

1. Yang, D.S.C. *et al.* *Nature* **333**, 232–237 (1988).
2. Raymond, J.A. & DeVries, A.L. *Proc. natn. Acad. Sci. U.S.A.* **74**, 2589–2593 (1977).
3. Ananthanarayanan, V.S. & Hew, C.L. *Biochim. biophys. Res. Commun.* **74**, 685–689 (1977).
4. Fletcher, N.H. *The Chemical Physics of Ice* (Cambridge University Press, 1970).
5. Hol, W.G.J., van Duijn, P.T. & Berendsen, H.J.C. *Nature* **273**, 443–446 (1978).

Roger H. Pain is Professor of Physical Biochemistry at the University of Newcastle upon Tyne, Newcastle upon Tyne NE2 4HH, UK.

Infrared astronomy

Glowing quasar cocoons

Daniel W. Weedman

DESPITE much intensive effort, astronomers do not yet understand how quasars, bright events at galactic nuclei, form and why they were more conspicuous in the early (distant) Universe than in the present (nearby) Universe. Ten ultra-luminous infrared galaxies studied by D.B. Sanders *et al.* (*Astrophys. J.* **325**, 74–91; 1988) show spectral characteristics similar to those of quasars and each is interacting gravitationally with another neighbouring galaxy. The authors believe that this interaction could be an important process in quasar evolution. A corollary is that quasars do not become visible at other wavelengths until the surrounding dust, which produces the infrared luminosity, can be dissipated.

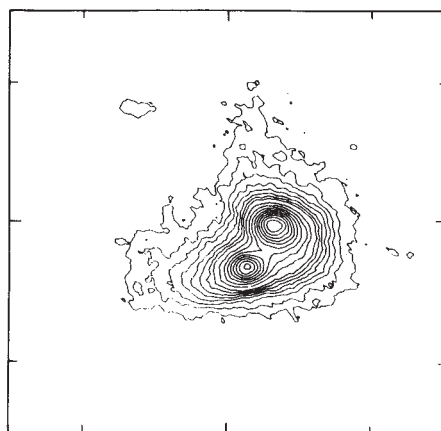
Because the Infrared Astronomical Satellite (IRAS) surveyed the entire sky in various infrared wavelengths, it has been possible to accumulate large and well-defined galaxy samples. From such samples, Sanders and collaborators picked the 10 having the greatest infrared luminosity, without any preselection as to the kind of galaxy involved. These are extraordinary galaxies, producing more than 10^{12} solar luminosities in the infrared portion of the spectrum (8–1,000 μ m). Once chosen using infrared luminosity as the sole criterion, these galaxies were observed with ground-based optical and millimetre-wave telescopes to elucidate their characteristics (see figure).

There are several unexpected similarities among all galaxies in the sample. All have strong emission line spectra at optical wavelengths, with the emission lines arising primarily in the galactic nuclei. Based on the widths and intensity ratios of these lines, nine of the ten are classifiable as having luminous, non-thermal sources of ionizing radiation in their centres, and therefore have active nuclei with spectral characteristics like those of quasars. In fact, two had already been classified years

before by optical observers as bright 'Seyfert' galaxies. Such galaxies with active nuclei have long been considered to be that subset of galaxies containing quasars in their centres.

More surprising are the consistent but unusual images of all galaxies in the ultra-luminous sample. Seven are not individual galaxies at all but are galaxy pairs, with the two galaxies close enough to be gravitationally interacting and distorting one another. The remaining three have nearby companions and show structural distortions that are evidence that interactions occurred in the recent past.

Finally, seven of the galaxies are also very luminous in CO molecular emission, indicating that they are rich in molecular gas. This is a highly consistent though not unexpected result. Previous observations of various samples of IRAS-selected galaxies have shown correlations between



Contour map from a visible light image of the ultraluminous infrared 'galaxy' IRAS08572+3915. The two cores in the image are two galaxies which are so nearly merged that they share a common envelope. The other nine galaxies studied by Sanders *et al.* also have peculiar structures, and all have quasar-like spectra. (From Sanders, D.B. *et al.* *Astrophys. J.* **325**, 74–91; 1988).

molecular and infrared emissions; these correlations are not too surprising because the infrared luminosity is thought to arise as black-body radiation from warm dust grains mixed within dense clouds of interstellar material.

Sanders *et al.* consider these observations to be overwhelming evidence that the phenomenon of ultraluminous infrared galaxies is a consequence of collisions between individual galaxies. Furthermore, given that these same galaxies have nuclear quasars, they argue that this process must be fundamentally related to the origin of quasars. They conclude that quasars are spawned by the collision of two gas-rich galaxies, and that the increased frequency of such collisions in a younger, denser Universe naturally explains why quasars were more conspicuous in the past.

These ideas are not completely new, as earlier observers had noted that a surprising number of quasars discovered optically have interacting galaxy companions. What is unique about the hypothesis put forward by the infrared astronomers is their contention that most quasars begin as ultraluminous infrared sources, and that optical observers cannot see them until the intense radiation has driven away the dust cocoon that shrouds the quasar birth. Sanders *et al.* make some preliminary suggestions as to where their individual galaxies lie in such an evolutionary sequence, and they are undertaking more observations to describe better the transition from infrared to optical quasar.

Now that it is clearly demonstrated that infrared-selected samples can reveal important aspects of quasars, fundamental questions must be addressed. What is the ultimate source of the energy heating the dust to make such ultraluminous infrared sources? How much of this energy arises within the non-thermal nucleus and how much within starbursts inside the molecular clouds around the nucleus? Is a quasar formed where none existed before as a consequence of the galaxy collision, or does a previously dormant black hole in a galaxy nucleus simply flare up as a new supply of gas is dumped into its vicinity? Were all early quasars so shrouded that optical surveys could never find the new quasars in the young (distant) Universe? And perhaps most puzzling, where does all of the dust come from that is necessary to explain the infrared radiation?

For the time being, further progress can only come using ground-based observations to target the galaxies found and sorted by IRAS. No further help from new space observations will come until the launch of the Infrared Space Observatory (ISO), expected in 1992. □

Daniel W. Weedman is in the Department of Astronomy, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

Seismology

The deep roots of continents

Thorne Lay

CONTINENTS are the Earth's most conspicuous structures; their precise role in the dynamic evolution of the planet, however, remains elusive. Early notions of continental drift, motivated by observations requiring large-scale translation of continents, were repeatedly stymied by the lack of a mechanism capable of propelling relatively thin (30 km) continental masses through or over the globally extensive oceanic crust. The subsequent theories of seafloor spreading and plate tectonics, which treat the oceanic plates as the upper thermal boundary layer of a mantle convection system, allow the embedded continental nuggets to drift as the boundary layer circulates. It is conventionally asserted that any thermal or compositional differences between the mantle underlying continents and that underlying oceans vanish below depths of 150–200 km. Lerner-Lam and Jordan¹, however, now provide strong evidence that relatively fast seismic velocities under stable continental regions extend more than 220 km into the mantle.

There is no question that continents are chemically differentiated, buoyant structures that resist recycling down into the mantle, and hence strongly influence the kinematics of surface plates; but the controversial assertion that the stable continental shields have roots extending as much as 400 km into the mantle has radical implications for the dynamic role of continents in plate tectonics.

Jordan² first advanced the thick-plate model for continents, basing it primarily on observations of seismic body and surface waves that he interpreted as evidence for fast seismic velocities underlying continents to depths of 400 km or more. For many years controversy surrounded this interpretation, and new analyses have only recently enhanced the resolution of the vertical extent of seismic heterogeneity.

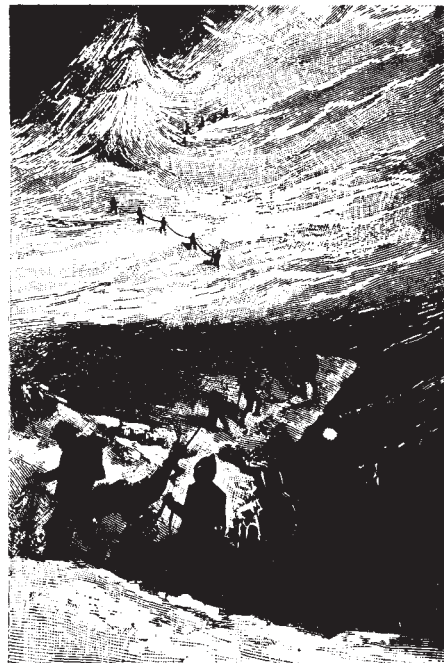
The best resolution of the upper-mantle seismic-velocity structure is obtained by analysing groups of seismic waves that bottom at different depths in the mantle, including waves that are multiply reflected off the surface. The waves can either be analysed as discrete body-wave arrivals^{3,4} or as interfering surface-wave overtones^{1,5,6}.

Lerner-Lam and Jordan in their new work¹ adopt the latter approach for fundamental and higher-mode 'Rayleigh' wave arrivals traversing either old oceanic or stable continental paths. The variety of Earth models obtainable from sparse seismic-wave data precludes an absolute determination of separate upper-mantle velocity structures under continents and

oceans. Thus the authors combine a hypothesis-testing procedure with the structure inversions to focus explicitly on the required depth extent of lateral variations between the two types of region. They find that lateral variations in the velocity structure of primary (P) waves can be confined to the uppermost 220 km of the mantle; lateral variations in the secondary- or shear-wave (S) velocity structure, however, must extend deeper, with satisfactory models having significant differences between oceanic and continental regions as deep as 400 km. Similar variations in S-velocity structure have been obtained in previous studies²⁻⁵, but the lack of any formal analysis of the resolution has always left the models' reliability uncertain.

Most seismic evidence seems to support the contention that differences of shear velocity between mantle beneath continents and that beneath oceans, extend more than 220 km into the upper mantle. But even having accepted this, the interpretation of this heterogeneity and its dynamic implications is still controversial. Anderson⁷ emphasizes the intrinsic ambiguity of interpreting relative velocity variations, arguing that the mantle more than 200 km under continental shields is normal unmolten mantle, whereas the hotter suboceanic mantle has anomalously low shear velocities because of partial

100 years ago



M. Richard's team proved that it was possible to live at high altitude on Mont Blanc. From *Nature* 38, 35; 10 May 1888.

melting and other isobaric phase changes.

Jordan^{2,8} instead attributes the high velocities below 200 km under shields to petrologically distinct material which accumulated during the early chemical fractionation of the continents. This compositional root must move with the underlying continental plate and it must resist the erosive effects of thermal convection in the surrounding system. It is also necessary to juggle possible chemical and temperature effects in the deep root to explain the absence of geoid (gravitational) anomalies associated with shields.

Additional evidence for the unusual properties of the deep continental environment have come from thermobarometry studies⁹ of mantle xenoliths (emplaced rock fragments) and modelling studies of surface heat-flow measurements within and along the margins of continental shields¹⁰. These indicate that low temperatures underlying the upper 200 km of continental shields have persisted for up to 3,500 million years, despite dynamic processes which tend to destabilize thick thermal boundary layers. A thick, durable chemical and thermal boundary layer which formed very early in the continental evolution could explain this. But the interpretation is complicated by the intrinsic tendency of continents to override cool subducting oceanic plates, which provides an alternative mechanism for maintaining long-term lower temperatures beneath continents⁷.

Although continental areas cover only one-third of the Earth's surface, and the stable shield regions are only a fraction of this area, it is still critical to resolve the thickness of the coherently translating surface plates. The maximum depth of the convective return flow in the mantle which accommodates the oceanic motions is not well resolved, and the possibility that continents with 400-km thick keels must be rafted about clearly affects our understanding of both the nature of the return flow and the forces that drive the motions.

How then, is this issue to be resolved? High-precision seismic investigations like that of Lerner-Lam and Jordan¹ must be undertaken to quantify the global occurrence of high-velocity extensions under shields. But this will not resolve the issues clouding the interpretation of the seismic models. It is clear that temperature-induced variations are important for lateral variations of seismic velocities

within continental regions. In fact, the most extreme lateral variations in shear velocity in the upper 300 km of the mantle that we know of are found between the Canadian shield and the tectonic belt of western North America. Temperature causes stronger variations in shear velocities than in compressional velocities, and thermally activated processes are also the cause of anelastic attenuation of seismic waves.

Thus, modelling of regional variations in seismic attenuation, together with studies of the elastic velocities, could be the key to apportioning the chemical and thermal interpretations of the deep continental root controversy. □

Thorne Lay is in the Department of Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA.

Transcriptional activation

Acid blobs and negative noodles

Paul B. Sigler

THE more that is known about the amino-acid sequences of proteins that participate in transcriptional activation, the clearer it becomes that many of the critical events cannot depend upon the precise geometrical complementarity that we associate with the interactions of globular proteins during molecular assembly, and the binding of substrates, cofactors and haptens. The latest entries in this chronicle of molecular imprecision are the activator domains of the proteins that stimulate

for correct placement and orientation with respect to the start site of transcription¹. The promoter sequences are a defined distance upstream of the transcriptional start site. Modulating the activity of the incipient transcriptional complexes is a wide variety of transcriptional activator proteins that bind to specific DNA targets called by various names, such as 'enhancers', upstream activating sequences (UASs) or hormone response elements (HREs). The position and polarity of

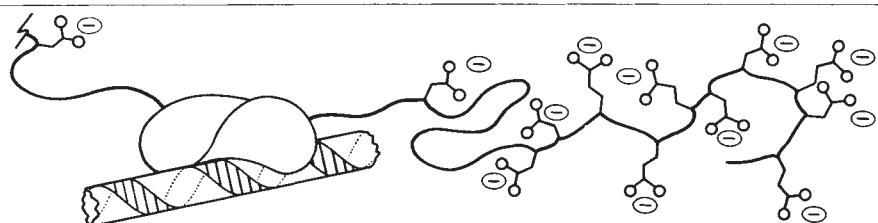


Fig. 1 A generic transcriptional-regulatory protein. A DNA-binding domain with a well-defined conformation (often dimeric) and a peptide with an excess of negatively charged side chains that has a poorly defined conformation (here, none at all).

transcription by RNA polymerase II. These activating 'structures' (and one must use that term advisedly) are targeted to specific DNA-sequences usually by a specific DNA-binding domain on the same polypeptide. The DNA-binding domains appear to have well-defined structural motifs; by contrast, mutational studies of the activator domains suggest a disquieting picture of a conformationally ill-defined polypeptide that can function almost irrespective of sequence, provided only that there is a sufficient excess of acidic residues clustered or peppered about. Thus, the elegant molecular machinery that so reliably controls expression of eukaryotic polymerase II-transcribed genes relies on nearly shapeless molecular components dubbed "acid blobs" or perhaps more aptly "negative noodles" (Fig. 1).

The function of these activating proteins and their poorly structured acidic domains is to stimulate the formation and/or action of transcriptional preinitiation complexes. The core of these complexes consists of the promoter-binding proteins on which the RNA polymerase depends

these *cis* regulatory sequences are not stringently defined with regard to the promoter that they activate. Thus in a geometrical sense, though not in a topological sense, the complete preinitiation complex must have a poorly defined architecture. Moreover, not only are the preinitiation complexes casually defined in spatial terms, but they have a surprisingly indefinite stoichiometry. It is not uncommon to find many enhancer sequences clustered or sprayed over a region of several hundred base pairs upstream or sometimes downstream of the promoter.

In most cases, reiterated sequences work better than just one or two. (Imagine haemoglobin benefiting from having a few extra β -subunits!) Even if we knew nothing about the sequences of the transcriptional activating proteins, this wide latitude of arrangement and stoichiometry argues strongly against any dependence of the activating events on specific and rigid (or even firm) complementary surfaces. One imagines the polymerase II transcription initiation complex before initiation as a convoluted loop of B DNA that brings the activator domains into contact with

1. Lerner-Lam, A.L. & Jordan, T.H. *J. geophys. Res.* **92**, 14007–14026 (1986).
2. Jordan, T.H. *Rev. Geophys. Space Phys.* **13**, 1–12 (1975).
3. Grand, S.P. & Helmlinger, D.V. *Geophys. J. R. astr. Soc.* **76**, 399–438 (1984).
4. Grand, S.P. *J. geophys. Res.* **92**, 14065–14090 (1987).
5. Woodhouse, J.H. & Dziewonski, A.M. *J. geophys. Res.* **89**, 5953–5986 (1984).
6. Tanimoto, T. *Geophys. J. R. astr. Soc.* **89**, 713–740 (1987).
7. Anderson, D.L. *J. geophys. Res.* **92**, 13968–13980 (1987).
8. Jordan, T.H. *Phil. Trans. R. Soc. A* **301**, 359–373 (1981).
9. Boyd, F.R. & Gurney, J.J. *Science* **232**, 472–477 (1986).
10. Ballard, S. & Pollack, H.N. *Earth planet. Sci. Lett.* **85**, 253–264 (1987).

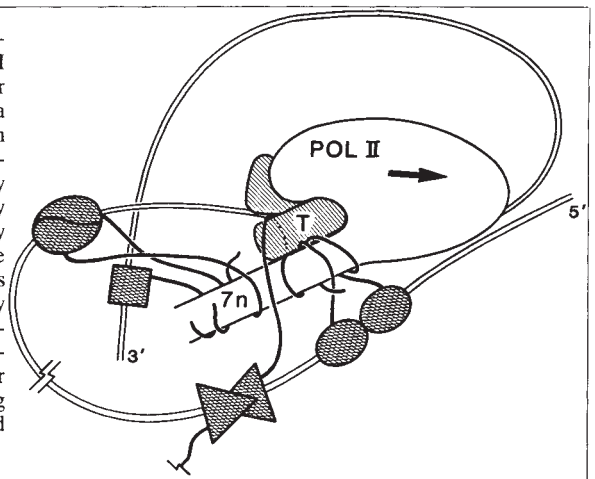
the polymerase and/or its promoter-binding proteins (Fig. 2).

The distinctly modular design of transcriptional activating proteins was first established in the study of the GAL4 and GCN4 activators of transcription in yeast. These proteins have domains that are remarkably independent from one another. Deletion of the activator leaves DNA-binding intact. When the activator domains of GAL4 and later GCN4 were fused to the bacterial DNA-binding protein, LexA, the chimaeric proteins enhanced transcription from promoter-reporter constructions downstream from LexA binding sites^{2,3}. Most transcriptional activating proteins appear to have this type of modularity; in fact, it is not even essential that the activator domain be covalently linked to the DNA-binding module. VP16, of relative molecular mass 65,000 (65K), is carried into the host by herpes simplex virus and is required to activate transcription of the viral immediate early genes. VP16 does not interact directly with the activating DNA sequences but with a cellular protein that does⁴. VP16 has an intensely acidic region within its 78 amino-acid carboxy-terminal segment that is required for activation and, when fused to the amino-terminal DNA-binding module of GAL4, strongly activates expression in yeast of the GAL4-responsive transcription units (S. L. McKnight, J. Ma and M. Ptashne, personal communication).

The GAL4-VP16 fusion indicates that virtually any acid blob or negative noodle will do, even one with a sequence that is radically different from the endogenous one and that is designed to function on viral enhancers in mammalian host cells. Recent experiments indicate that this phylogenetic generality works both ways. Both Ptashne's and Chambon's groups have reported that GAL4 or GCN4 will stimulate transcription from their cognate DNA-binding sites when these sites are introduced upstream from viral promoters in mammalian cells^{5,6}. Moreover, these alien transcriptional-activator systems from yeast can work synergistically with endogenous mammalian counterparts such as the simian virus 40 enhancer or the system that mediates the transcription response to steroid hormones.

The most convincing display of sequence-indifference in the acidic peptide was reported by Ma and Ptashne⁷, who found that moderate transcriptional activation could be achieved by fusing the GAL4 DNA-binding domain to random polypeptides encoded by fragments of genomic *Escherichia coli* DNA. Activating sequences were found to contain an excess of acidic side chains. As in the mutational studies of the GAL4 activating domain⁸, the strength of activation is correlated roughly with the preponderance of

Fig. 2 A hypothetical preinitiation complex of polymerase II (POL II), shown with a globular catalytic domain (arrow) and a CT7n appendage (7n) to which the negative noodles of the transcriptional activators strongly adhere in a conformationally indefinite fashion. The strongly crosshatched domains bind the activator proteins to enhancers on the DNA loops. The lightly crosshatched structures are proteins that associate with polymerase II and/or its promoter elements (T, TATA binding factor) and can also be contacted by negative noodles.



negative charge. Although it is possible that the various acidic activator domains require small, possibly transient, secondary structural elements⁹, activating sequences of this type clearly do not form a compact globular domain with a distinct tertiary structure.

What kind of assembly could activate polymerase II transcription using an unspecified number of ill-structured negative noodles positioned on DNA elements contiguous with but far from the site of action? The key may lie in the structure of eukaryotic RNA polymerase II itself. In 1985, two groups independently reported that the carboxy-terminal domain in the largest subunit of yeast and mouse RNA polymerase II was composed of a 27- and a 52-heptapeptide repeat, respectively^{10,11}. Except for one variable position, the repeating block of the yeast and mouse domains is virtually identical (Fig. 3). More recent studies in *Drosophila* and other eukaryotes indicate that the carboxy-terminal repeating 7-mer (I will



Fig. 3 Canonical sequence of CT7n. Because it repeats, this represents one of seven possible circular permutations. The residues underlined the most are the most invariant.

call it CT7n) is a general feature of eukaryotic RNA polymerase II. The exactness of this striking repeat suggests a structurally well-defined domain dominated by a unique secondary structure. Except for the proline residues that presumably help to stabilize the secondary structure, nearly all the amino-acid side chains of the repeating blocks are hydrophilic (primarily hydroxyl groups and almost never acidic side chains), suggesting that this domain projects out from the globular fold of the remaining polypeptide and is fully exposed to solvent (Fig. 2). This picture is consistent with the observation that endogenous proteases readily cleave the repeating elements from the catalytic domain.

There is suggestive evidence of a regulatory role for CT7n. The repeating

element is not needed for accurate transcription in a minimal system, indicating that the catalytic function resides in the amino-terminal domain and that the carboxy-terminal repeating elements are not needed for accurate polarized positioning on the promoter¹². Deletion experiments indicate that, depending on the size of the original CT7n domain, a 'stub' of 11 (yeast) to 23 (mammalian) repeating elements is required for cellular viability and that an intact CT7n domain supports cellular metabolism best¹³. As with the negative noodles of the activator regions, exact size, sequence and cellular origin are not critical: the larger CT7n of the hamster can function in yeast when fused to its endogenous counterpart¹⁴. Prokaryotic RNA polymerase is very similar to the amino-terminal eukaryotic catalytic domain, but does not have a CT7n domain and exhibits little regulation by distant elements. One can imagine that the presumably rigid secondary structure of CT7n forms a kind of hitching post bristling with hydroxyl groups to which a negative noodle could bind (either alone or in concert with an intermediate protein) through strong but varied hydrogen-bonded interactions.

Figure 2 depicts a possible model of the preinitiation complex. It is based on the supposition that negative noodles of the transcriptional activating proteins attach through their DNA-binding domains to the appropriate *cis*-activating sequences to form an assembly with polymerase II at the designated promoter¹⁴. The stabilizing interactions are between the carboxylates of the noodle and the hydroxyl groups of CT7n (Fig. 4a). The promoter-binding proteins doubtless potentiate some of these interactions and may also interact directly with certain *trans* activators. Because of the richness of potential hydrogen-bonded contacts, global complementarity of precisely interacting surfaces would be neither necessary nor desirable, because the CT7n domain of a given organism must serve a wide variety of different transcriptional-activation systems. ►

Once the components of a stable pre-initiation complex have been recruited and assembled, the polymerase must be released from the complex to proceed with transcription. The model also suggests how this may occur. The hydroxyl groups of CT7n could be phosphorylated at or near the interaction sites of the activator domains (Fig. 4b). The flexibility of the noodles and their variable patterns of attachment would allow a kinase molecule access to the abundant hydroxyl groups of the CT7n. Indeed, J. Corden and S. McKnight (personal communication) suggest that acidic residues might actually activate the phosphorylation. Insinuating negatively charged phosphate groups would repel the negatively charged noodle

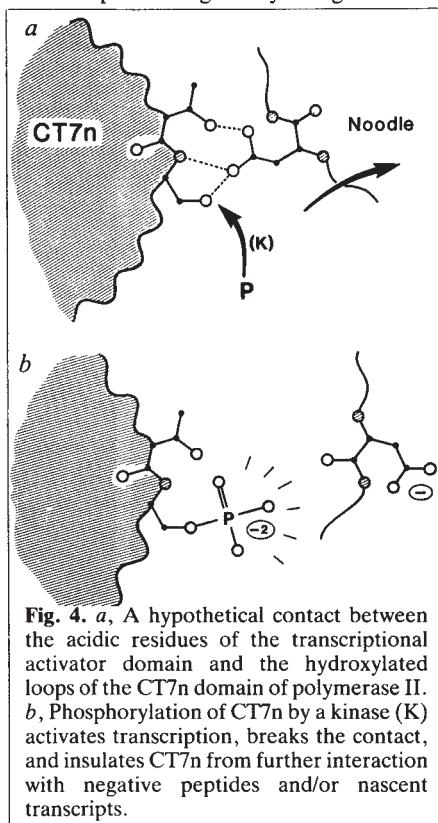


Fig. 4. *a*, A hypothetical contact between the acidic residues of the transcriptional activator domain and the hydroxylated loops of the CT7n domain of polymerase II. *b*, Phosphorylation of CT7n by a kinase (K) activates transcription, breaks the contact, and insulates CT7n from further interaction with negative peptides and/or nascent transcripts.

and possibly other inhibiting factors. Moreover, phosphorylation would prevent potential 'pausing' or 'braking' interactions caused by hydrogen bonds between the hydroxylated CT7n and the newly synthesized transcript. Thus, phosphorylation would disassemble the pre-initiation complex while activating the transcriptional function of the enzyme.

The model has been left vague about the role of the important TATA-binding factor. Evidence reviewed at a recent meeting* suggests that the binding of this class of proteins to the TATA element of certain promoters is augmented by remotely bound *trans* activators (R. Roeder, Rockefeller University). However, if the TATA factor is a globular protein of modest size, it is difficult to imagine

how it could make significant direct contact with a large and indefinite number of diverse structures. Difficulties of this sort generally reflect insufficient information. Until more is known of its structure and function, the exact role of the TATA factor(s) will remain elusive. Promising biochemical and genetic studies of the yeast TATA factor(s) by S. Hahn in Guarente's laboratory (MIT) should provide valuable insights.

The activation of polymerase II transcription is not the only case in which extended polypeptides of ill-defined tertiary structure mediate important interactions. Exported or integral membrane proteins have similarly loosely defined sequences that specify their destination and orientation; the unstructured basic amino-terminal arms of the viral capsid subunits help anchor the protein to the underlying nucleic acid.

Cro, λ and *trp* repressors have flexible carboxy- and amino-terminal arms that contribute to DNA affinity. All these systems share with the transcriptional initiation complex the need for a mechanism by which many and various proteins can interact with a common cellular element. These flexible and variable contact patterns depart from the traditional view of specific molecular interactions gained from studying assemblies of globular molecules that give crystalline images. Whereas crystal structures at atomic resolution are crucial to our understanding of the physical chemistry and dynamics of specific molecular interactions, we can imagine many assemblies like the eukaryotic transcriptional initiation complexes whose function requires strong but less precisely defined arrangements than the ones we have seen crystallographically. The model presented here will certainly be wrong in many details, but it typifies the need for a less traditional view of molecular assemblies and new types of structural experiments to close the gap in our chemical understanding of cellular regulatory systems. □

1. Struhl, K. *Cell* **49**, 295-297 (1987).
2. Brent, R. & Ptashne, M. *Cell* **43**, 729-736 (1985).
3. Hope, I. A. & Struhl, K. *Cell* **46**, 885-894 (1986).
4. Triezenberg, S. J., Kingsbury, R. C. & McKnight, S. L. *Genes Dev.* (in the press).
5. Kakidani, H. & Ptashne, M. *Cell* **52**, 161-167 (1987).
6. Webster, N., Jin, J. R., Green, S., Hollis, M. & Chambon, P. *Cell* **52**, 168-179 (1987).
7. Ma, J. & Ptashne, M. *Cell* **48**, 847-853 (1987).
8. Gill, G. & Ptashne, M. *Cell* **51**, 121-126 (1987).
9. Giniger, E. & Ptashne, M. *Nature* **330**, 670-692 (1987).
10. Corden, J. L., Cadena, D. L., Ahearn, J. M. & Dahmus, M. E. *Proc. natn. Acad. Sci. U.S.A.* **82**, 7934-7938 (1985).
11. Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. *Cell* **42**, 599-610 (1985).
12. Zehring, W. A., Lee, J. M., Weeks, J. R., Jokerst, R. S. & Greenleaf, A. L. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
13. Bartholomei, M. S. *Molec. Cell. Biol.* **8**, 330-339 (1988).
14. Allison, L. A., Wong, J. K.-C., Fitzpatrick, V. D., Moyle, M. & Ingles, C. J. *Molec. Cell. Biol.* **8**, 321-329 (1988).
15. Cadena, D. L. & Dahmus, M. E. *J. biol. Chem.* **262**, 12468-12474 (1987).

Paul B. Sigler is in the Department of Biochemistry and Molecular Biology at the University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA.

Daedalus

Cash in hand

THE cashless society, in which all transactions are carried out by inserting a plastic card into a computer terminal, is slow in arriving. One reason is its great vulnerability to theft and misuse of the cards. Now Daedalus proposes an unstealable cash-card: the user's own thumbnail.

A nail has a smooth and uniform surface. A small focused laser moving over it in a dot-matrix could easily emboss small white thermal decomposition marks in the body of the nail (such marks sometimes occur naturally). A bit larger than the dots on a compact disc, they will be easily optically readable, but shielded from chance abrasion. A nail has no nerves, so laser-writing will be quite painless.

Every day a nail grows about 0.1 mm, enough to record perhaps ten new transactions; the user of the system will accumulate on his thumbnail a running financial statement which will act in effect as a continually updated password. Each time he inserts his thumb into a terminal to validate a transaction, the system will check that statement against its files. If everything matches, it accepts the new transaction and prints it below the previous ones. But if there is a suspicious mismatch, the terminal sounds the alarm and clamps down on the suspect thumb, trapping the fraudster till the police arrive.

Thus fraud is powerfully discouraged. The only feasible way to cheat the system might be for a crooked manicurist to photograph a client's thumbnail, and use it to make a forged thumb. But by the time it was ready, the victim would probably have used the system again, so the forgery would not bear the latest data. It would be detected, and trapped by the terminal into the bargain, for the police to study later as evidence.

The system also imposes a certain financial prudence on its users. The spendthrift who fills his thumb up with wild transactions will soon be choked off by sheer lack of space. But diseases that arrest nail-growth (like mumps and measles) will rapidly threaten bankruptcy. And wearing one's finances on one's thumbs will have its own risks, for secrecy will be hard to maintain. Gigolos with magnifiers may revive the old gallantry of hand-kissing; palmists will scan both sides of your hand; business executives of both sexes may take to wearing opaque nail-varnish, and start consigning their nail-clippings to a microshredder. And a death in the family will have the worried relatives hastily falsifying the deceased's thumbs with a laser scrambler, before some unscrupulous undertaker or body-snatcher can copy or remove them in order to steal their encoded legacy.

David Jones

* The UCLA Symposium DNA-Protein Interactions in Transcription was held in Keystone, Colorado on 4-10 April 1988.

Are Trojans pursuing the eclipsing pulsar?

SIR—In this issue, Fruchter *et al.*¹ report the discovery of a pulsar in an eclipsing binary system. The pulsar is moving in a nearly circular orbit (eccentricity < 0.001). The orbital parameters suggest that the companion has a very low mass, much lower than the probable masses of the companions of other known radio and X-ray binary pulsars. Interestingly, the mass is sufficiently low that the Lagrange points L_4 and L_5 of the restricted three-body problem are points of stable equilibrium. There is thus the possibility that smaller agglomerations of matter are orbiting the system near these points. The situation could be analogous to that of the Trojan asteroids, which oscillate about L_4 and L_5 in the Sun–Jupiter system.

The pulsar mass function is²

$$f = \frac{M_1 q^3 \sin^3 i}{(1 + q)^2} = 5.20 \times 10^{-6} M_\odot \quad (1)$$

where M_1 is the pulsar mass, i is the orbital inclination angle and $q = M_2/M_1$, where M_2 is the companion mass. The Lagrange points are stable provided $q \leq q_{\text{crit}} = 0.0385$ (see, for example, ref. 2). From equation (1), this will be true provided $\sin i \geq \sin i_{\text{crit}} = 0.412 (M_1/1.4M_\odot)^{-1/3}$. For a pulsar of mass $1.4M_\odot$, this constraint requires $i > i_{\text{crit}} = 24^\circ$, which is almost certainly guaranteed by the presence of an eclipse. The dependence of this result on the assumed pulsar mass is very weak. Only a pulsar of mass $\leq 0.1M_\odot$ could be eclipsed and yet have unstable Lagrange points. (Note that the minimum mass for a neutron star is $0.1M_\odot$ (see, for example, ref. 3).) In fact, for the likely values $\sin i = 1$, $M_1 = 1.4M_\odot$, the companion mass is $M_2 = 0.022M_\odot$ and the mass ratio q is $1/64$, well within the stability range for the Lagrange points.

The eclipse is centred on orbital phase 0.25, when the companion is between the pulsar and the Earth. The orbital phases of L_4 and L_5 are at $0.25 \pm \theta/2\pi$, where $\theta = \tan^{-1}[\sqrt{3}(1+q)/(1-q)]$.

Because the range of q for stable Lagrange points is small, the corresponding range in phase is small. As q varies from 0 to q_{crit} , the phase of one Lagrange point varies from 0.42 to 0.44, while the phase of the other varies from 0.08 to 0.06. The detection of matter at these phases would confirm that the mass ratio is below q_{crit} . Because of the weak dependence on q , however, it is unlikely that a precise mass ratio could be determined. Moreover, matter would probably librate about the

Lagrange points. Were it possible to track the orbit of such libration, or even to measure the period, the mass ratio could then be determined².

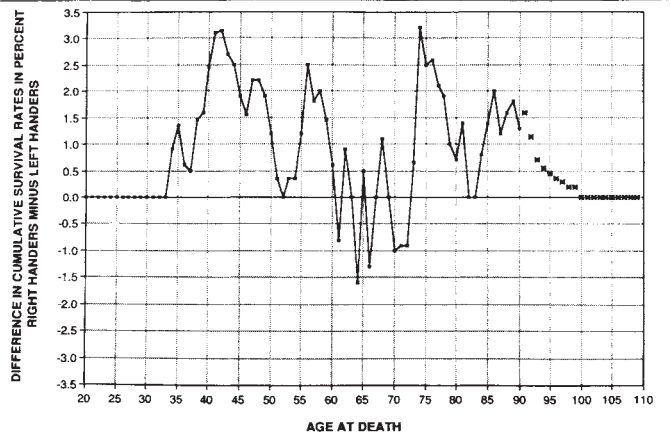
This binary system is most probably still very active. There may be a powerful beam of high-energy particles from the pulsar and a strong stellar wind from the companion¹. Thus it is very possible that nothing will be found at either of the Lagrange points, even if they are stable.

Do right-handers live longer?

SIR—Several findings suggest that left-handedness may be associated with reduced longevity. For instance, Porac and Coren¹ reported that 13 per cent of 20-year-olds are left-handed but only 5 per cent of those in their fifties and virtually nobody of 80 or above. We believe that this absence of left-handers from the oldest age groups reflects higher biological and environmental risk.

To investigate further the relationship between handedness and age of death,

The relative survival rate of right- and left-handers. Each point represents the difference between the cumulative percentage survival of the two groups. Points above the line, higher survival rates for right-handers; points below the line, higher survival rates for left-handers. Stars, ages at which there were no surviving left-handers.



we have analysed all baseball players listed in *The Baseball Encyclopedia*² for whom dates of birth and death, as well as throwing and batting hand, are reported. A subject was assigned to a handedness group only when both throwing and batting hand were the same with no indicated change in hand use.

Mean age at death for the 1,472 right-handers was 64.64 years (s.d. = 15.5) and 236 left-handers was 63.97 years (s.d. = 15.4). This difference is difficult to interpret as the range is so large and the distribution is skewed. However, a nonparametric test of group differences (Wald–Wolfowitz runs test) indicated that the greater longevity for right-handers is significant ($Z=6.63$, $P<0.001$). To clarify the pattern of results, we also examined data for the cumulative proportion of individuals surviving at each age. We found that the groups are virtually identical in mortality until the age of 33. From

On the other hand, this is the first system outside the Solar System where the conditions are favourable for observing matter at stable Lagrange points.

STUART L. SHAPIRO
SAUL A. TEUKOLSKY

Center for Radiophysics and Space
Research and
Departments of Astronomy and
Physics,
Cornell University, Ithaca,
New York 14853, USA

that age onwards about 2 per cent more right-handers than left-handers survive at each age ($P<0.001$) (see figure).

Although these data suggest that left-handedness is associated with reduced longevity, they do not provide any direct causal information. Three factors seem possible. First if prenatal and perinatal birth stressors are more probable in left-handers^{1,3–6} it may reduce their ability to survive. Second, genetic effects and intra-uterine hormones may have reduced the effectiveness of the immune system of sinistrals, increasing the likelihood of an earlier death⁷. Finally, left-handers may

have more accidents in an environment designed for a right-handed majority¹. These factors may act singly or in concert to account for the fact that left-handers seem to have a shorter life span.

DIANE F. HALPERN
Department of Psychology,
California State University,
San Bernardino,
California 92407, USA

STANLEY COREN
Department of Psychology,
University of British Columbia,
Vancouver V6T 1Y7, Canada

1. Fruchter, A.S., Stinebring, D.R. & Taylor, J.H. *Nature* **333**, 237–239 (1988).
2. Danby, J.M.A. *Fundamentals of Celestial Mechanics* (Macmillan, New York, 1962).
3. Shapiro, S.L. & Teukolsky, S.A. *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects*, Ch.9 (Wiley, New York, 1983).

1. Porac, C. & Coren, S. *Lateral Preferences and Human Behavior* (Springer, New York, 1981).
2. Reichler, J.L. (ed.) *The Baseball Encyclopedia* (Macmillan, New York, 1979).
3. Bakan, P. *Nature* **229**, 195 (1971).
4. Coren, S. & Porac, C. *Behav. Genet.* **10**, 123–138 (1980).
5. Searleman, A., Porac, C. & Coren, S. *Can. J. Psychol.* **36**, 478 (1982).
6. O'Callaghan, M.J. *et al. Lancet* **i**, 1155, (1987).
7. Geschwind, N. & Galaburda, A.M. *Cerebral Lateralization: Biological Mechanisms, Associations, and Pathology* (MIT, Cambridge, Massachusetts 1987).

It is also important to recognize that selection at the level of the species is not predicted upon the punctuational model. Species selection simply looms larger in this model than in a gradualistic framework, which grants gradual evolution a larger role in the determination of macro-evolutionary trends. In fact, although he

never analysed the process in detail, Wright⁶ asserted that selection operates at the level of the species, and even Darwin's *Origin of Species* alludes to this process.

STEVEN M. STANLEY

Department of Earth and
Planetary Sciences,
The Johns Hopkins University,
Baltimore, Maryland 21218, USA

1. Maynard Smith, J. *Nature* **330**, 516 (1987).
2. Maynard Smith, J. *Nature* **332**, 311–312 (1988).
3. Stanley, S.M. *Macroevolution* (Freeman, San Francisco, 1979).
4. Stanley, S.M. *Evolution* **36**, 460–473 (1982).
5. Stanley, S.M. *Proc. natn. Acad. Sci. U.S.A.* **72**, 646 (1975).
6. Wright, S. *Am. Nat.* **90**, 5–24 (1956).

Unusual segregation of cystic fibrosis allele to males

SIR—One in fifty chromosomes seven in caucasian populations has the cystic fibrosis (CF) mutant gene. A strong linkage disequilibrium has recently been described with the DNA probes KM-19, CS-7 and XV-2c^{1–3}, suggesting that most CF chromosomes in most European populations are derived from a single mutation event. It is not clear whether the mutation expanded through some form of carrier advantage, or by founder effect followed by genetic drift. It has been suggested that the high frequency of the CF gene could have arisen through increased fertility of CF carriers^{4,5}; recently Pritchard⁶ reported a slight but unexpected increase of children in the families of uncles of CF children. He also reported an increase in the proportion of male children in these families.

To investigate the significance of that observation in a definitive way we have undertaken the haplotyping of asymptomatic CF carriers by DNA analysis. We investigated 50 French families, each with a CF individual, in which there was also at least one normal sibling, to determine whether the sibling is homozygous for the unaffected allele, or a CF carrier. Among the 60 siblings investigated, who include about the same number of boys and girls, we observe a striking distortion of segregation with sex. Twenty of 22 normal homozygotes (91%) are girls ($\chi^2 = 18.1$, $P < 0.0001$), and 16 of 21 (76%) heterozygotes carrying the paternal CF chromosome are boys ($\chi^2 = 5.76$, $P < 0.02$). In contrast, the maternal CF chromosome has been passed equally to boys and girls.

This unexpected result arises from the fact that CF chromosomes of paternal origin, unlike the normal chromosomes, are passed preferentially to males. Note that, as previously reported, the overall proportion of carriers to homozygous normals is ~ 2 to 1, as expected³.

We hope that others will test families in this way to determine whether this result is statistical chance, or can be confirmed in other populations. If generally true, it suggests a mechanism by which the CF gene pool can be expanded through preferential inheritance of the affected chromosomes from male to male. Such a mechanism could involve close linkage of CF to some determinant of sperm viability or fitness, or selection during early embryonic development. It also suggests that the proportion of CF carriers will be higher in males than in females.

A. KITZIS

J. C. CHOMEL

J. C. KAPLAN

Institut de Pathologie Moléculaire,
CHU Cochin, 75014 Paris, France

G. GIRAUD

A. LABBE

B. DASTUGUE

CHU Clermont-Ferrand, France

V. DUMUR

J. P. FARRIAUX

PH. ROUSSEL

CHU Lille, France

R. WILLIAMSON

St Mary's Hospital Medical School,
London W2 1PG, UK

J. FEINGOLD

INSERM U 155, Paris, France

1. Estivill, X. *et al.* *Nature* **326**, 840–845 (1987).
2. Estivill, X. *et al.* *Genomics* **1**, 257–263 (1987).
3. Beaudet, A. L. *New Engl. J. Med.* **318**, 50–51 (1988).
4. Danks, D. M., Allan, J. & Anderson, C. M. *Ann. hum. Genet.* **28**, 323–335 (1965).
5. Knudson, A. G. Jr, Wayne, L. & Hallett, C. M. *Am. J. hum. Genet.* **19**, 388–392 (1967).
6. Pritchard, D. J. *Nature* **330**, 319 (1987).

'Cytoplasmic' and 'maternal' inheritance

SIR—Altenberg's attempt (*Nature* **331**, 309; 1988) to clear up the confusion between the use of 'maternal inheritance' and 'maternal effect' is helpful but still leaves unnecessary confusion. There is agreement that when the genotype of the maternal parent determines the phenotype of the progeny regardless of the progeny's genotype (as in the classic example of shell coiling in *Lymnaea peregra* snails discussed by John Galloway *Nature* **330**, 204; 1987), the appropriate term is 'maternal effect'.

Confusion is introduced by using the term 'maternal inheritance' for a phenomenon which should more appropriately be called 'cytoplasmic inheritance'. In higher organisms, where the egg cell is the only source of extranuclear hereditary factors, the term 'maternal inheritance' is accurate. The term 'cytoplasmic inheri-

tance' is more appropriate because it also describes the identical phenomenon in organisms such as protozoa where use of 'maternal' would not be correct, and it avoids the linguistic confusion of 'maternal effect' and 'maternal inheritance'.

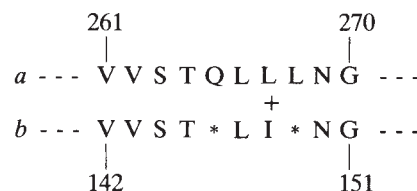
This is subtly recognized in the dictionary definition of maternal inheritance which indicates that the phenomenon is also called cytoplasmic inheritance, although there is no analogous qualification under the definition of 'cytoplasmic inheritance' (R.C. King & W.D. Stansfield, *A Dictionary of Genetics*; Oxford University Press, 1985). The use of the term 'maternal inheritance' is an unnecessary and confusing representation of the underlying biological basis of the inheritance pattern being described.

BRUCE HAGGARD

Hendrix College,
Conway, Arkansas 72032, USA

HIV and HLA similarity

SIR—Recent data suggest a functional role for a conserved region of the human immunodeficiency virus (HIV) envelope glycoprotein gp120: mutations introduced into this sequence¹ and antibodies directed against the concerned region² block virus infectivity without affecting virus binding to the CD4 receptor. I would like to point out that this region has a striking resemblance to a 10-amino-acid region in the extracellular, membrane-proximal domain of HLA class II β -chains (see figure).



Comparison of HIV gp120 (a) and the consensus HLA-DR, -DP and -DQ β -chain (b) sequences. Stars, dissimilar residues. Plus sign, a conservative change.

HLA class II molecules and gp120 are both recognized by membrane components of CD4-positive T lymphocytes. In both cases^{1,3}, the 10-amino-acid region is distant in the primary amino-acid sequence from the predicted CD4-binding sites. It is conceivable that this short sequence contributes to both antigen presentation and virus infection by interacting with an as yet undefined component of the T-cell membrane, distinct from the CD4 molecule. If so, HIV could have evolved to exploit more than one T-cell component of the antigen presentation machinery to infect its target cells.

JOHN A.T. YOUNG

University of California Medical School,
San Francisco, California 94143, USA

1. Willey, R.L. *et al.* *J. Virol.* **62**, 139–147 (1988).
2. Ho, D.D. *et al.* *Science* **239**, 1021–1023 (1988).
3. Golding, H. *et al.* *Nature* **317**, 425–427 (1985).

Scientific Correspondence

Scientific Correspondence is intended to provide a forum in which readers may raise points of a scientific character. They need not arise out of anything published in *Nature*. In any case, priority will be given to letters of less than 500 words and five references. □

Padé approximation and linear prediction methods

SIR—Yeremian and Claverie¹ use a method combining the Padé rational approximation and the Laplace transform to extract the coefficients and exponents in multi-exponential data. They emphasized three important features: (1) The assumption of the number of decaying exponential components involved in the total signal is not necessary; (2) increased numerical accuracy results if the Taylor series of the Laplace integral is made in the neighbourhood of the actual decay rates; and (3) their method can solve the Lanczos multiple exponential problem. We wish to point out that the above features were already known and could be implemented in the framework of Prony's linear prediction² (LP) and z-transform³ theory. More importantly, the LP methods using the Householder decomposition⁴ (LPQRD) or the singular value decomposition⁵ (LPSVD) are more stable and accurate than the Longman method⁶ used by them. Both LPQRD and LPSVD methods have been used for some time to extrapolate truncated signals and analyse signals of exponentially damped sinusoid without *a priori* assumption of the total number of spectral components. The functional form of the Padé approximation is very similar to the linear prediction z-transform (LPZ) spectral formula³ that we have derived assuming the signal is composed from a finite number of damped sinusoids.

Unlike QRD or SVD, their Longman algorithm⁶ may have two limitations: (1) In the Longman approach, the LP coefficients in the Padé table are completely determined from the first $2M$ data points, where M corresponds to the LP filter length; and (2) The Longman algorithm can be unstable, especially when the first few data points are close to zero. On the other hand, QRD or SVD is favoured because of four points: (1) LPQRD or LPSVD is not subject to signal truncation problems. In contrast, the Laplace integral transform used in ref. 1 can be inaccurate if the signal is severely truncated because the Laplace integral should be evaluated in time from zero to infinity. (2) The decomposition methods work well with singular matrices and have better stability. (3) All data can be used to calculate the LP coefficients, yielding better accuracy than the Longman method which uses fewer data. And (4) QRD or SVD provides eigenvalues and matrix rank determination that can be used to reject noise components effectively. Once the LP coefficients are determined from any of the above methods, the frequencies and the decay rates can be evaluated by solving the roots of the polynomial formed by the LP co-efficients. Finally, the corresponding amplitude and phase for each spectral component can be determined from a linear least-squares fit.

To illustrate the advantages of the

LPQRD method over the Longman method, we present a comparison using a simulated signal of two close components (see table). Fifty data points are generated with a step of 0.1 s, where data values are rounded to the second decimal.

	Frequency (Hz)	Decay time (s)	Amplitude	Phase (degree)
Theoretical values	1.00	2.00	1.00	0.0
LPQRD	1.10	1.00	2.00	0.0
$M=5$	1.01	2.47	0.85	-16.4
	1.10	1.08	2.19	5.8
$M=10$	1.00	1.95	1.04	0.7
	1.10	1.00	1.96	-0.4
$M=20$	1.00	1.95	1.06	-0.5
	1.10	1.00	1.94	0.2
Longman's				
$M=5$	1.05	1.24	2.78	-0.3
	1.33	0.46	0.28	-4.1
$M=10$	1.01	1.62	1.47	4.0
	1.12	0.93	1.55	-3.5
$M=20$	1.01	2.21	0.95	6.1
	1.10	1.07	2.09	-13.5

In this test we compared both methods with a z-transform expansion at infinity, as in the conventional Prony method. They all failed to resolve two components with $M=3$ (not listed in table), and the Longman method still failed with $M=5$.

As nicely illustrated by Yeremian and Claverie¹, in many cases better accuracy can be obtained if the Taylor expansion of the Laplace integral can be made at some points other than with a z_0 at infinity. Similarly, within the framework of the LP and z-transform³ theory, the matrix equation of the linear prediction can be slightly

modified so that the LP coefficients are evaluated with z-transform in the neighbourhoods of the actual roots. The improvement of the resolving power and numerical accuracy in either their Laplace transform or our z-transform depends on how closely one can estimate the expansion centre or the actual roots. In the Padé method the Taylor series expansion of the Laplace integral can be made only at one particular point. Thus, clusters close to the expansion centre may be better resolved, but distant clusters may still remain unresolved. In contrast, the LP approach is preferable because the z-transform can be evaluated simultaneously around several clusters, resulting in a more global improvement in resolution. Finally, the Laplace integral transform is similar to z-transform, except that the former is in a continuous and the latter is in a discrete form. As most experimental data are recorded as discrete points, the z-transform seems to be more appropriate. Comparing the stability between the Longman method and the matrix decomposition method with a least-squares fit, the LP routine using QRD or SVD appears to be the method of choice.

J. TANG
J.R. NORRIS

Chemistry Division,
Argonne National Laboratory,
Argonne, Illinois 60439, USA

1. Yeremian, E. & Claverie, P. *Nature* **326**, 169–174 (1987).
2. Prony, R. *J. Ecole Polytechnique* **1**, 24–76 (1795).
3. Tang, J. & Norris, J.R. *J. chem. Phys.* **84**, 5210–5211 (1986).
4. Tang, J., Lin, C.P., Bowman, M.K. & Norris, J.R. *J. magn. Reson.* **62**, 167–171 (1985).
5. Barkhuijsen, H., De Beer, J., Bovee, W.M.M.J. & Van Ormondt, D. *J. magn. Reson.* **61**, 465–481 (1985).
6. Longman, I.M. *Int. J. Computer Math.* **B3**, 53–64 (1971).

How fecundity balances mortality in birds

SIR—Birth rates must balance death rates if natural populations are to persist. Saether¹ demonstrated that two components of fecundity, clutch size and age at maturity, help compensate for differences in mortality among European bird species. But fecundity also depends on a third variable, the number of clutches per year, for which data are available² on Saether's 107 species. Following Saether, we performed analyses at the family level ($n=34$). Number of clutches per year decreases with increased survival rate ($r=-0.634$, $P<0.001$, and thus also compensates for mortality. Annual fecundity, the product of clutch size and number of clutches, is even more closely correlated with survival ($r=-0.824$, $P<0.001$).

Saether¹ also considered the relationship of clutch size and age at maturity to each other and to survival after the effects of body size had been removed by partial correlation analysis. He found that families with larger clutches also have earlier ages at maturity when body weight is controlled for. We find that number of clutches per year is not significantly correlated with

clutch size ($r=0.143$) but is negatively correlated with age of maturity ($r=-0.478$, $P<0.01$) and with survival ($r=-0.374$, $P<0.05$) when body weight is partialled out. But clutch size is not significantly correlated with body weight across the sample of European birds ($r=-0.284$), whereas both number of clutches per year and age at first breeding are significantly correlated with body weight ($r=-0.658, 0.614$, both $P<0.001$).

Similar patterns of covariation among the fecundity factors and body weight occur in a much larger sample of 3,142 species from 149 families². We agree with Saether that the cause of such differences may be evolutionary or a direct result of density-dependent processes.

PETER M. BENNETT

Institute of Zoology,
Zoological Society of London,
London NW1 4RY, UK

PAUL H. HARVEY

Department of Zoology,
University of Oxford, Oxford OX1 3PS, UK

1. Saether, B.-E. *Nature* **331**, 616–617 (1988).
2. Bennett, P. M. thesis, Univ. Sussex (1986).

Back to Berggasse 19

John C. Marshall

Freud: A Life for Our Time. By Peter Gay. Dent, London/Norton, New York:1988. Pp.810. £16.95, \$25.

TO SAY that no biography of Sigmund Schlomo Freud will ever replace that by Ernest Jones is rather like making an analogous remark about the four gospels. In both cases, the biographers had unparalleled first-hand knowledge of their respective subjects, but (it has been argued) they were more concerned with promoting a myth than reporting the facts. One difference, however, is that Freud himself assiduously aided the promulgation of his legend; in the case of Jesus ben Joseph one might reasonably suspect that the subject would have been horrified at most of what was later said and done in his name.

Publication of an up-to-date, scholarly biography of Freud, one that takes full account of the historical research (and speculation) that post-dates the completion of Jones's work in the mid-1950s should thus be quite an event. To what extent does Peter Gay's *Freud* (with its deliciously ambiguous subtitle *A Life for Our Time*) live up to our expectations?

Anyone in search of new evidence about the soap-opera aspects of Freud's life and times will be mightily disappointed. What happened to the mysterious second wife of Jacob, Freud's father? Did Freud have an affair with Minna, his wife's sister? Or — a more sinister accusation — was Freud's colleague Max Eitingen a KGB agent? Gay (correctly) concludes that the material currently available will not support the inverted pyramids of speculation that have recently been erected on these topics. Nonetheless, as a responsible historian, he is forced to cover himself against the possibility that emulators of such as Jeffrey Masson and Peter Swales may succeed in prying the relevant evidence loose. The real scandal of Freud scholarship lies in "the restrictive policy of Freud's guardians". As Gay remarks, the practice of "either denying or slowing down access to important materials" can only nourish rumours.

How, then, does Gay's account differ from Jones's? The most striking overall impression lies in the extent to which the narrative structure of the earlier book has been toned down. Jones unashamedly constructed a hero's life, as pure and

unblemished as the most favourable reading of the documentary evidence would allow. A revelation of the truth isolates our hero from his blinkered peers; a small group of disciples slowly forms; one by one they betray the master; Freud fights on alone until the world is finally convinced of his profound insights into the human psyche.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Freud — "never ceases to fascinate".

By contrast, Gay sets the story much more firmly within a coherent social and scientific context; he can admit that Freud's behaviour towards his colleagues (and patients) was not always beyond reproach, and that the evidence he adduced in support of his theories was sometimes "peculiar". Freud was never as isolated or as generally reviled (even in his native Vienna) as Jones claimed, and Gay presents a far more balanced picture of Freud's relationship to contemporary medical circles.

With regard to psychoanalytical politics, Gay is somewhat less convincing. He has a tendency to deflate criticism of Freud's overweening ego by quoting the founder's own highly humorous version of events. Thus Freud writes to Lou Andreas-Salomé: "I have never fought against differences of opinion within the circle of *ψα* research, especially since I usually have more than one opinion about one issue — that is, before I publish one of them". Anyone who has engaged in any form of intellectual pursuit will react sympathetically to those lines . . . although it may not have been quite so easy for Freud's

immediate colleagues to accept that they too had to change their opinions the instant the master pronounced.

Yet one must be grateful to Gay for painting a picture of Freud that is human, all too human. And never more so than in the portrayal of Freud's relationship — passionate, tender and protective — with his beloved daughter Anna. Gay relates this story in greater detail than any previous biographer and only a parent with a heart of stone could fail to be deeply moved by it.

In other respects, Gay's own regard for Freud — as man and scientist — is more open to criticism. Take, for example, the notorious experimentation with cocaine. Gay describes Freud's role in prescribing

large amounts of cocaine for his friend Fleischl-Marxow as "ill-advised" when a somewhat stronger term would have been appropriate. Concerning Freud's own intake of cocaine, Gay writes that "he continued to use it in modest quantities at least until the mid-1890s". How "modest"? We do not know. But we do know that on 12 June 1895 Freud wrote to Wilhelm Fliess "I need a lot of cocaine". The issue is of some importance in the light of E.M. Thornton's attempt to interpret Freud's own medical and psychological history as a case-study in the effects of severe cocaine addiction. In a bibliographical appendix, Gay dismisses Thornton's book *Freud and Cocaine: The Freudian*

Fallacy as "a model in the literature of denigration". It is indeed that, but the claims therein may nonetheless be true. For Gay to refuse to take the *possibility* seriously is hardly an adequate riposte.

If Freud did formulate his theories under the influence of cocaine, that would not, in itself, militate against their validity. A more pertinent objection is that Freud (and his colleague Breuer) misdiagnosed their early, supposedly 'hysterical' patients. On this topic, Gay is very close to silent; he simply ignores the substantial achievements of Thornton's book, and of such papers as Hurst's "What was wrong with Anna O?" (published in the *Journal of the Royal Society of Medicine* in 1982). These authors put forward a strong *prima facie* case that the patients on whom psychoanalysis was founded may have suffered from such conditions as tuberculous meningitis, temporal lobe epilepsy, disseminated encephalomyelitis or Tourette's syndrome. The evidence is, of course, circumstantial, and it is always hazardous to propose such retrospective neurological diagnoses on the basis of case-reports that are resolutely psycho-

Mary Evans

analytical. Yet Gay can hardly object to the endeavour when he himself is prepared to accept Leonardo da Vinci as one of Freud's analysands! If Gay is unwilling to confront these issues directly he cannot claim to have written a life of Freud for our time.

The problem is in fact far more widespread than the question of failure to make an accurate organic diagnosis. Gay's book, with its skilful interweaving of Freud's life and work, is least successful at precisely that point when Freud the

moments Freud claimed merely that analysis might transform neurotic despair into ordinary unhappiness.

Gay is loath to question Freud's status as a scientist. When he was still primarily a clinical neurologist, Freud was unquestionably a scientist of the first rank; when he became an analyst, doubts arise. The separation of neurology from psychiatry that took place in the early parts of this century (and for which Freudian theory was partly responsible) may yet prove to have been an unmitigated disaster. No argument in Gay's text prompted me to reconsider my long-held view of Freud *qua* analyst: Freud was the first great novelist to chronicle the spiritual decline of the Austro-Hungarian Empire.

It is in this sense that Freud's life is indeed "a life for our time". The man never ceases to fascinate and Peter Gay's biography holds the attention compulsively throughout its 810 pages. As a precondition of being allowed to leave 'Greater Germany', Freud was required to sign a statement to the effect that he departed of his own free will, and that he and his family had been well treated. As befits a resident in the city of Karl Kraus

and Kurt Waldheim, Freud was no stranger to the uses of irony. He appended to the document the laconic remark: "Ich kann die Gestapo jedermann auf das beste empfehlen" (I can most highly recommend the Gestapo to everyone). No man with the chutzpa to write such a vitz in these circumstances can be all bad. But for a man who tried to live without illusions, one can only be thankful that he never knew the fate of his sisters, Rosa, Marie, Adolfine and Pauline. Freud attempted in vain to rescue "the four old women" from the coming holocaust; they all died in the concentration camps. □

John C. Marshall is in the Neuropsychology Unit, part of the Neuroscience Group at the Radcliffe Infirmary, Oxford OX2 6HE, UK.

• Freud indeed "never ceases to fascinate" — publishers, at least. Shortly to be issued by Basil Blackwell is Christopher Badcock's *Essential Freud* (hbk £22.50, \$24.95; pbk £6.95), a book for the general reader as well as psychologists, behavioural scientists and so on. From Yale University Press comes *Freud in Exile: Psychoanalysis and its Vicissitudes*, edited by Edward Timms and Naomi Segal, published on 6 June to mark the 50th anniversary of Freud's arrival in London. Price is £25, \$35.



From Freud

The entrance to Berggasse 19, Freud's home from 1891 until 1938 when the family left Austria after the Anschluss.

analyst begins to supersede Freud the neurologist. Little slips start to appear in the text. Thus, contrary to Gay's assertion, John Hughlings Jackson was a neurologist; he was no more (and no less) a psychologist than was every other major behavioural neurologist of the time. When Freud, in his great work on aphasia (1891), extended the range of behavioural observations to which a theory of the aphasia should be responsive he was not denying that the proximate cause of such pathologies was a structural brain lesion. Why, then, did he fail to take seriously the possibility that his 'hysterical' patients manifested some symptoms that were a direct consequence of organic pathology? And that other 'symptoms' may reflect the patient's own attempts to make sense of his or her impaired cognitive functioning? Is there any good reason to believe that psychoanalytical interpretations are more helpful in allowing patients to come to terms with their organic disorder than frank discussion of the likely consequences of structural damage would be? The 'talking therapy' will not cure organic pathologies, but in his more sober

Names and claims

Anthony G.E. Pearce

The Paraneuron. By T. Fujita, T. Kanno and S. Kobayashi. Springer-Verlag: 1988. Pp.367. DM 348, £124, \$210.90.

I DO know what a paraneuron is, but if you don't, and unless you have a relish for propaganda, I recommend that you proceed with caution. Written by two anatomists and a physiologist, *The Paraneuron* has, in truth, a single begetter (Fujita) and a single purpose.

Technically the book is a first-class production, with remarkably few typographical errors. Copiously illustrated throughout with excellent light and electron micrographs, virtually every one of its 30 chapters features also an extravagance of black or blue and black line diagrams. These latter, no doubt, are designed to simplify, for students, the often complex relationships between individual paraneurons and their immediate neighbours. Additional simplification is achieved by the provision, at the end of each chapter, of a summary of its content.

The work is divided into two parts, general and special. The first 11 chapters, which constitute Part 1, run a gamut from the eclectic history of the paraneuron concept, via stimulus-secretion coupling, through production, transport and effects of messenger substances — all paraneurons are purinergic (p.63) — to ontogeny, kinetics and neoplasia. Part 1 ends with a

discourse on phylogeny in which proto-stomian aspects overwhelm the interests of the deuterostomia.

Throughout this part, and indeed throughout the whole book, the reader is subjected to a barrage of neologisms which stem from the recognition, by the senior author, of the benefits which accrue from the successful coining of a label (GEP, for example). But the authors' treatment of the paraneologism "paracrinia" is hardly fair to its originator Friedrich Feyrter (1895–1973) in that it repeats Fujita's somewhat arrogant opinion, voiced in 1983, that "neither Feyrter himself nor recent advocates of his concept have precisely defined" the adjective paracrine (the word paracrinia is never used). The authors proceed to take credit for remedying this supposed omission. But in a letter to *The Lancet* (1976), written as a friendly answer to the assertion by the late Morton Grossman that Feyrter's "para" should be equated with "additional", I quoted two out of many more expressions by which Feyrter clearly defined paracrine as "acting solely on immediate [*anliegenden*] neighbours or, broadly speaking, as acting locally [*an Ort und Stelle*]" . At once graciously accepted by Grossman, this interpretation has been current since that time, except perhaps in Japan.

The worst of Fujita's neologisms, "paraneuromas" is the title of Chapter 10. The authors are revealed here as totally out of touch with modern oncology (not excluding its linguistic derivation). Hastening, too late, to ridicule use of

the historical appellation apudomas for tumours of the neuroendocrine system, they appear to be unaware that the accepted terminology for these is now neuroendocrinoma and neuroendocrine carcinoma.

Part 2 of the work has 19 chapters, each embracing a single cell type or a family of related cells. There are thus too many cells and too many chapters for pertinent comment here. Coverage of each cell type sensibly follows a more or less uniform pattern: distribution, products, innervation, electrophysiology, ontogeny and phylogeny. In Chapter 13, curiously, the thyroid parafollicular cells are referred to as "C (from clear) cells", instead of the usual and correct view that "C" stands for Cope's calcitonin! But the oddest inclusion is the subject of Chapter 27, the mast cell. Here its accepted mesenchymal origin is emphasized as typifying the authors' view (p.85) on the derivation of paraneurons from all the germ layers.

Having come this far, there may still be some readers inclined to ask why it was necessary to write this book at all, except as an illustrated text for students of anatomy. Yet perusal of the preface alone provides some insight. Its variety of hopeful expressions, and its oblique references to the works of Starling, Lyell and Darwin, clearly anticipate the forthcoming descent of the mantle of these great men upon the authors of *The Paraneuron*. Those who read further will have no trouble in detecting the vested interest of this reviewer, whose published contributions to neuroendocrinology are referred to selectively and frequently, though skilful handling of relevant literature is not restricted to the work of any one author or group of authors. With respect to my own contributions, the paper quoted is invariably that bearing the latest possible date of publication. Thus is effectively elided the chronological distance between the APUD concept (the neuroendocrine system) and its paraneuronal equivalent (see *Nova acta Leopoldina* 56, 118–124, NF255; 1984).

Composed as it is, with the benefit of hindsight, it is necessary to conclude that *The Paraneuron* is neither more nor less than a gigantic take-over bid for the equity of that large section of neurobiology which has hitherto been called neuroendocrinology. Obviously I am not prepared to recommend the book to those unversed in this field; but for those who are, it makes interesting reading.

The choice between neuroendocrinology and paraneuronics is a clear one, and I imagine that the distinction will be maintained without difficulty. □

Anthony G.E. Pearce is Emeritus Professor of Histochemistry in the University of London and Consultant Pathologist to Hammersmith Hospital (the Royal Postgraduate Medical School), Du Cane Road, London W12 0HS, UK.

Stellar systems in theory

Frank Shu

Galactic Dynamics. By James Binney and Scott Tremaine. Princeton University Press: 1988. Pp. 733. Hbk \$75, £42.10; pbk \$25, £14.

SINCE 1970 or so, study of the structure, dynamics and evolution of stellar systems such as galaxies and star clusters has undergone a renaissance. Books on the subject have become badly outdated, and a new treatment has long been overdue. The new graduate text by James Binney and Scott Tremaine, two of the world's leading practitioners in this field, fills the gap admirably. The manuscript circulated for a few years while undergoing preparation and revision, and among the *cognoscenti* it has already achieved the status of a classic. Now it is available as an attractive volume from Princeton University Press; the resultant product has been well worth the wait.

The book begins with an observational overview, and then dives deeply, with a minimum of splash, into authoritative discussions of virtually the entire range of topics currently being worked on: potential theory (mass models, that is, density-potential pairs), orbits and integrals of motion, dynamics of collisionless systems (equilibria and stability), spiral structure, stellar encounters, dynamical friction and mergers, theory of star clusters (evaporation, core collapse and the effects of binaries), galactic evolution (chemical as well as in colour and luminosity) and the role of dark matter in the Universe. Except for the last two chapters, the emphasis throughout is heavily theoretical, with relatively little cross-referencing to the companion observational volume *Galactic Astronomy* written by Mihalas and Binney (published by W. H. Freeman). Nevertheless, the discussions are well motivated from the physical and astronomical points of view, and they are self-contained, the more technical manipulations being banished to a series of mathematical appendices. For material too complex to present in a textbook format, readers are referred to the original papers in the research literature, after having been given an introduction at a level that has adequately prepared them to profit from such reading.

The strengths of the book are many. In discussing the work of others, the authors have not been content merely to reproduce the original arguments, but have reworked them to make the exposition more transparent or closer to the developments in the rest of the book. For example, their derivation of Toomre's sequence of mass

models for galactic disks, starting with a simple physical visualization of Kuzmin's disk, is a model of elegance and brevity. Binney and Tremaine also take potentially complex topics — such as the phase-space structure of regular and irregular orbits, the equilibria of triaxial galaxies, the cannibalism of a small galaxy by a larger host, or the evolution of globular clusters — and reduce them to their dynamical essentials without losing the flavour of the richness of the physical phenomenon. Their handling of more incompletely understood problems, such as the nature of and evidence for dark matter, is equally concise, while remaining well balanced, up to date and full of insight.

The book does have its eccentricities. For example, the method of characteristics is never introduced *per se* as a standard technique for solving quasilinear partial differential equations, of which the 'collisionless' Boltzmann equation is a prime example. Thus, the authors' demonstration of the mathematical relationship between the constants of motion derived from orbit analyses and the most general solution for the distribution function of an encounterless stellar system shows sufficiency, but not necessity, and deprives the reader of a useful tool of the trade. Also, although Binney and Tremaine generally do an excellent job of presenting all sides of controversial topics, their judgement is occasionally coloured in favour of work that has been carried out at Princeton and the two Cambridges (places where they have spent appreciable time). For example, the acerbic comment on p. 607 that "The mass-to-light ratio of binary galaxies is probably large, but not so large as the ratio of the masses of papers on this subject to the light they have shed on it" does a disservice to a serious subject. Fortunately, such slips of taste are rare, and the authors usually make intelligent and positive use of the opportunity, expressed in their preface, "to advertise [their] own views on the proper organization and relative importance of various topics and areas".

For a comprehensive and penetrating treatment of modern problems of the structure and evolution of stellar systems, this book has no peers. I strongly suspect that copies of *Galactic Dynamics* will find permanent homes not only in libraries, but in the offices and studies of all serious students of the field. □

Frank Shu is a Professor in the Department of Astronomy, University of California, Berkeley, California 94720, USA.

• The latest volume in Cambridge University Press's astrophysics series is *X-ray Emissions from Clusters of Galaxies*, by Craig L. Sarazin. The book deals largely with the observational properties of X-ray clusters of galaxies and with the theory of the physical state, dynamics and origin of the hot intracluster gas. Price is £32.50, \$49.50.

Principle approach

Chris Quigg

Longing for the Harmonies: Themes and Variations from Modern Physics. By Frank Wilczek and Betsy Devine. Norton: 1988. Pp. 361. \$19.95.

SCIENCE'S great gift to human understanding of the material world can be epitomized in a single word — simplicity. When applied to descriptions of reality, simplicity conveys two interrelated meanings, economy of expression and economy of hypotheses, or laws. Experience has taught us that the most useful descriptions of nature can be formulated precisely and elegantly. We do not believe that a separate agent lies behind each phenomenon in the physical Universe, but have come to idealize the notion, which we can trace to William of Ockham, that a few laws that apply in the same fashion at all times and in all places should be sufficient to describe reality.

How are we to arrive at such laws? A style of thought that has continually been rewarded is to take seriously great principles, to follow their consequences to the logical conclusion and not to sacrifice them lightly. By taking seriously Newton's theory of gravity, Adams and LeVerrier inferred the existence of the planet Neptune from anomalies in the orbit of Uranus. By taking seriously the law of conservation of energy, Pauli conjectured the neutrino from the energy spectrum of electrons emitted in radioactive beta-decay.

In such cases, the new invention is not offered as a replacement for existing law, but as a logical outcome that must be true, if existing law is still to hold. The operative question is not, "Can you disprove this?", but "How could it be otherwise?"

What John Archibald Wheeler has called the "radically conservative" strategy is the unifying theme of this elegant book by Frank Wilczek, one of the most imaginative and wide-ranging of theoretical physicists, and engineer-writer Betsy Devine. Wilczek and Devine characterize radical conservatism as "conservative in its reluctance to introduce new assumptions. . . . Creative tension and power is added to it by a radical approach to the few assumptions that are adopted. These assumptions must be formulated precisely and pushed as hard as possible. Their consequences must be fully drawn; they must be applied to as many situations as possible in the natural world and to the oddest and extremest conditions we can set up in the laboratory".

Uniformity, a second pervasive theme, is introduced by showing that we and the stars are made of the same stuff, so we can learn in terrestrial laboratories the physical

laws that govern the structure of the Universe, and we can apply here on Earth what we learn by studying the heavens. The puzzles of uniformity — why are all the fundamental building blocks the same throughout the Universe, and why is the Universe so uniform — are raised, and our current understanding explained.

In between, many important topics in the quest to understand the microscopic structure of matter and the large-scale structure of the Universe are introduced in turn. A discussion of the cosmic distance ladder leads to Hubble's discovery of the expansion of the Universe, and to the Big Bang. The wave-particle duality of quantum mechanics is developed with care, and the combination of relativity and quantum mechanics is shown to lead ineluctably to antimatter and virtual particles. The elements of what has come to be called the Standard Model of Elementary Particle Physics are developed in a logical and insightful manner. Pauli's principle, which accounted for the periodic table of the elements, and the notion of quarks lead by the radically conservative path to the necessity of a charge of the strong interaction called 'colour', from which it is irresistible to formulate the promising and comprehensive theory of the strong interactions, quantum chromodynamics. The notion of hidden symmetries is the basis for our current understanding of the weak interaction, and prompts us to imagine more perfect worlds, in which the full symmetry might be manifest. There the dream of a unified theory of the fundamental interactions would be reality.

Throughout the book, the science is sound and well chosen, the writing is graceful and the analogies are apt. I particularly enjoyed the autobiographical interlude, "How Asymptotic Freedom Discovered Me", for its description of the historical setting in which Wilczek and his mentor David Gross undertook the calculations that led them to discover (at the same time as David Politzer) that in gauge theories the effective coupling strength could decrease at short distances or high energies. It was great theoretical sport in the early 1970s to show that observations of deeply inelastic scattering of electrons from protons could not be reconciled with relativistic quantum field theory. Some players took pleasure in proving that field theory itself must be wrong; others derived amusement from showing that the model of the proton as a collection of quasi-free quarks ('partons'), enunciated by Richard Feynman and others to interpret the experiments, was inconsistent with quantum field theory and therefore had to be incorrect. As Wilczek relates it (and it coincides with my recollection as a spectator), the goal of the Gross-Wilczek work was to drive the last nail into the field-theory coffin by studying gauge theories, the class of field theories left for last because of the techni-

cal difficulty of the calculations. Their plans were dashed when they found that in gauge theories quarks behaved as if they were free of the strong interaction, at asymptotically high energies. They had not found what they were looking for, but something more interesting than they might have dared to dream. Wilczek's account is a refreshing counter-example to the belief that scientists always know where they are going.

Longing for the Harmonies is a fine account for the general reader of how science is done, and of the interplay between theory and experiment that is crucial to progress. It gives an up-to-date picture of our struggle to understand the Universe as a whole, its basic constituents and the interactions among them. It is not, however, a complete picture of the physical world. The same themes of uniformity and the radically conservative strategy apply, with equal success, to what lies in between the very large and the very small, and to phenomena associated with collections of simple systems. Wilczek and Devine correctly note that "It is no small part of the charm of physics that, after more than three hundred years of astonishing progress, there remain absolutely major and fundamental problems". It is no less a part of the appeal of our science that it has a great and growing unity. Dare we hope for a sequel in which the phenomena that operate on the scale of everyday experience, of condensed matter and complex systems, are linked with those of the cosmos and the microworld? □

Chris Quigg is Deputy Director for Operations of the Superconducting Super Collider Central Design Group, Lawrence Berkeley Laboratory, Berkeley, California 94720, USA, and Visiting Professor in the Department of Physics, University of California at Berkeley.

New in paperback

- *Calcium in Muscle Activation* by Johann C. Rüegg, in a corrected second printing. Publisher is Springer-Verlag, price is £28. For review see *Nature* 325, 584 (1987).
- *Animal Architecture and Building Behaviour* by Michael H. Hansell. Publisher is Longman, price is £9.95. For review see *Nature* 311, 183 (1984).
- *Sweet Track to Glastonbury: The Somerset Levels in Prehistory* by Bryony and John Coles. Publisher is Thames & Hudson, price is £12.95. For review see *Nature* 323, 120 (1986).
- *Spinors and Space-time, Vol. 2: Spinor and Twistor Methods in Space-time Geometry* by R. Penrose and W. Rindler. Publisher is Cambridge University Press, price is £17.50, \$34.50. For review see *Nature* 322, 318 (1986).
- *Introduction to Supersymmetry* by Peter G. O. Freund. Publisher is Cambridge University Press, price is £8.95, \$14.95. For review see *Nature* 326, 254 (1987).
- *The Launching of Modern American Science, 1846-1876* by Robert V. Bruce. Publisher is Cornell University Press, price is \$12.95 (United States), \$14.30 (elsewhere). For review see *Nature* 329, 209 (1987).

Simulating the climatic effects of nuclear war

Stephen H. Schneider & Starley L. Thompson

National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80307, USA

The climatic effects of a hypothetical large nuclear war have been simulated by an increasingly comprehensive series of global numerical models. Short-term climatic effects are now found to be less severe than predicted by early studies but the chronic long-term atmospheric effects remain potentially serious. The sum of all indirect effects could exceed those of blast and radioactivity.

THE possibility that nuclear war could have serious atmospheric after-effects was clearly demonstrated in the calculations by Crutzen and Birks¹ in 1982. They considered that fires started in the wake of a large nuclear war could loft enough smoke into the atmosphere to obscure sunlight across a whole hemisphere for several weeks or more, with a consequent darkening of the Earth's surface and possible climatic response as well. Assessments of the scale of the effect by climatic simulation have been marked by scientific²⁻⁵ and political controversy⁶⁻⁹. The first estimate¹⁰ predicted average surface temperatures below freezing in land in the northern hemisphere, and included maximum decreases in surface temperature of ~35 °C. Because this is the difference in temperature between summer and winter in mid-latitudes Richard Turco, the principal author of this study, coined the phrase 'nuclear winter' to describe the effect.

The result came from a one-dimensional model known as TTAPS, from the initials of its authors, in which climate was averaged annually over a land-covered planet and the response to massive injections of smoke into the atmosphere was simulated. In the past five years, several generations of climatic models of increasing complexity have been used, all of which have confirmed the general contentions of TTAPS, that injection of hundreds of millions of tons of smoke deep into the atmosphere may lead to substantial surface cooling. The most recently published calculations^{11,12}, however, predict a cooling of only about 10 °C in July and generally less in other seasons. As this is closer to the difference in temperature between summer and autumn, we have argued that for consideration of temperature effects alone, the label 'nuclear fall (autumn)' is more appropriate than nuclear winter as a description of the current simulations⁹. This re-characterization of the problem has been controversial¹³⁻¹⁵ and has prompted the reconvening of an international assessment group¹⁶. Because of the heated nature of some of the debates and the rapid evolution of the relevant scientific studies, we intend to review recent progress in simulating the climatic consequences of nuclear war, without dwelling on the earliest simulations. Nor are we able to cite all relevant recent work. Rather, we present a limited selection of representative recent results, focusing on our own research. The debates concerning policy implications including our own views may be found elsewhere^{6-9,13,17,18}.

Components

The overall problem of estimating the environmental consequences of nuclear war goes well beyond climate modelling. It depends on a sequence of issues, some marked by untestable uncertainties. Fundamentally, the entire assessment rests on the human behavioural assumption of a plausible scenario of nuclear attacks. Indeed, not only must the number of weapons detonated be known, but also the geographic aiming points, the altitude of burst, the size of the weapons, and even the duration and season of the war. Because these are, of course, impossible to know in detail, a sequence of plausible war scenarios have been constructed^{10,19-21}. Next, given some assumptions of local

meteorological conditions for atmospheric transmissivity estimates, calculations of the thermal flux radiated by the fireballs produced by nuclear explosions can be used to estimate the area of fire ignition for specific targets¹⁹. Estimates of smoke production require assumptions about the amount and kinds of available combustible materials (fuel) and the degree of blast-caused fragmentation of the fuel. These factors determine the rate and type of combustion, for example smouldering or flaming, and the efficiency with which smoke is produced per mass of fuel burned. The type of fuel and the manner in which it is burned also has a large effect on the fraction of elemental carbon (soot) in the smoke—the component that is critical for absorbing solar energy. Estimates of smoke and soot yield per unit mass of fuel are available from laboratory, field and forest fire experiments, and have been used by many to estimate smoke production amounts^{10,19,20,22,23}.

To estimate how much smoke is injected into an area a few hundred kilometres across (typical global atmospheric model's grid resolution), it is first necessary to estimate how fire-plume motions loft the smoke from individual targets, how much water vapour is injected by combustion and entrainment from low altitudes, and whether a convective 'capping cloud' and precipitation develop²⁴. These factors control the processes that could promptly remove smoke particles from the atmosphere, or otherwise change their number concentration, morphology or composition²⁰. Once dissociated from the fire area, individual plumes composed of sunlight-absorbing smoke, having widths and lengths of the order of tens to hundreds of kilometres, respectively, could then interact with the atmosphere or each other to create mesoscale (50–500 km) convective circulations that could further loft, remove or spread smoke²⁵. Only at this point is a climate model 'source term' available that can be used in global general circulation model (GCM) simulations of the atmospheric implications of the smoke generated by nuclear war.

A committee report of the US National Research Council (NRC)¹⁹ estimated that a baseline 6,500-MT war would produce 360 Tg (10¹² g) of smoke containing about 65–70 Tg of soot. In the NRC report, however, only half the smoke was assumed to survive prompt removal in convective capping clouds, so it was estimated that 180 Tg of smoke containing ~35 Tg of carbon would be injected into the atmosphere. A detailed estimate (R. Small, unpublished abstract, Defense Nuclear Agency Global Effects Program Technical Review, 1988) of smoke produced in a 3,000-MT attack against the United States, if doubled to account for smoke from the Soviet Union and Europe, gives 78 Tg of smoke containing 54 Tg of carbon. This recent estimate, which does not account for any prompt removal, is therefore somewhat smaller than, but still consistent with, the NRC report estimates of 65–70 Tg of soot carbon produced. Penner²⁶ has summarized the uncertainties in smoke-production estimates and finds that the plausible range of smoke production for a large nuclear war lies between 24 and 400 Tg, with the corresponding carbon soot production lying between 10 and 225 Tg. Within these ranges, there is a large uncertainty about the most

probable values. These ranges encompass the three hypothetical smoke scenarios (approximately 15, 50 and 150 Tg carbon) recently chosen by the ENUWAR committee of the Scientific Committee on Problems in the Environment as simulation test cases¹⁶. Most climate modellers have used soot amounts of ~ 50 Tg^{9,12,27-29} as baseline smoke injections for a large nuclear war. Depending on assumed optical properties, this amount of soot, or equivalent amount of smoke, would produce a visible absorption optical depth of ~ 1.0 – 1.5 if spread evenly over a hemisphere. Optical depth is a measure of the degree to which the smoke prevents sunlight passing through it. For example, for solar-absorbing aerosol of unit optical depth, a solar beam impinging on the atmosphere at a 60° angle from the zenith would be attenuated by a factor of $\exp[-\sec 60^\circ \times 1] = 0.14$.

Of course, it is not the post-war atmospheric effects, but rather the biological consequences of those effects that are of greatest interest to people^{30,31}. Although we will not review them here, consideration of biological response has led climate modellers to examine not only mean temperature response at continental scales, but to consider the variability and extremes of biologically significant factors such as temperature, precipitation and sunlight reaching the surface.

Natural analogues

Because the results of modelling the after-effects of nuclear war are unverifiable by direct experimentation, some have suggested that certain natural events might provide quantitative insight. The most widely cited is the hypothesized collision of a comet or asteroid with the Earth some 65 million years ago at the Cretaceous/Tertiary (K/T) boundary³², a time coincident with mass extinctions of species on land and sea. The dust aerosol generated by the impact of a 10-km-diameter body has been estimated to be 5×10^6 Tg spread fairly uniformly around the globe for a period of several months, resulting in sub-freezing temperatures on all land masses and a global loss of photosynthesis for up to a year³³. Although there is some dispute over the geological evidence for such an impact³⁴, even if one took place, it is not a very useful analogy for smoke injection from nuclear war. Nuclear smoke would almost certainly not be uniformly spread about the globe in optically thick concentrations at the outset, but would rather be in dense patches, thinning out to a semi-transparent hemispheric or global pall in several months. The hemispheric optical depths for smoke would be more than four orders of magnitude less than has been estimated for the 'asteroid winter'. This scale difference combined with large quantitative uncertainties surrounding the impact/extinction theory make it difficult to derive any direct insights into the effects of nuclear war.

On the other hand, research on the nuclear winter hypothesis could have some direct bearing on the K/T extinction problem. Some have suggested that an elemental carbon layer found in rocks at the K/T boundary suggests that the impact fireball ignited widespread, simultaneous fires by radiative heating which created a global pall of smoke that led to extinctions³⁵. We believe, however, that it is much more likely that such a carbon layer represents smoke injected over a period of several years from the episodic burning of vegetation killed by the environmental effects of the impact.

Another oft-mentioned analogue is the aerosol injected into the atmosphere by volcanic eruptions, especially the Tambora eruption in 1815 which was followed by the legendary "year without a summer". However, this particular 'volcano winter' analogy bears little relation to nuclear winter¹³ as it is now reasonably well documented that the average northern hemisphere volcano-induced cooling in the summer of 1816 was only ~ 1 – 2°C , and that the sporadic frosts in New England and Europe that summer resulted from major distortions to the normal jet stream patterns³⁶⁻³⁸. Unless it can be demonstrated that the aerosol from the volcanic eruption induced large shifts in the atmospheric planetary-wave patterns, the argument that

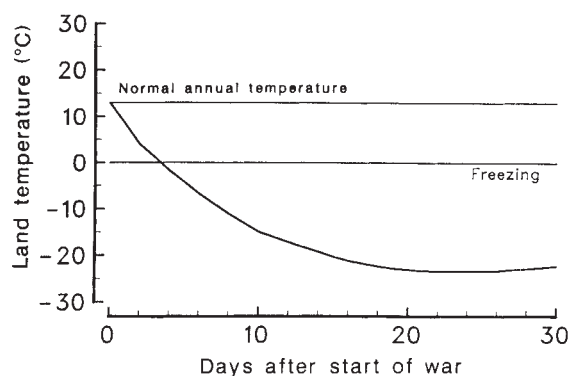


Fig. 1 The hemispheric, annual average surface temperature for a completely land-covered planet from the TTAPS¹⁰ baseline 5,000-MT case.

a cooling of a few degrees causes frosts in mid-latitude summer is unconvincing. Moreover, volcanic aerosols, in general, make poor analogues for the effect of smoke from nuclear war because the mechanisms of injection into the atmosphere, the morphology and chemical properties, and the optical properties of carbonaceous smoke versus volcanic ash or sulphuric acid droplet 'dust' are vastly different. We believe, however, that studies of the climatic effects of volcanic aerosols are useful for testing the sensitivity of models to the presence of aerosols of low optical depth, but are only modestly useful for the verification of some components (such as the stratospheric transport of aerosols) of the nuclear winter hypothesis.

The most directly applicable analogue is very large forest fires. Although such fires are relatively limited with respect to intensity and total smoke production compared to the sum of urban fires hypothesized to result from a large nuclear war, they can nevertheless give insight into the microphysical processes and long-range transport of smoke, and help to test the regional sensitivity of a climate model.

Three generations of climate models

Although Crutzen and Birks¹ speculated that the smoke from a nuclear war could be of global climate significance, the TTAPS study¹⁰ was the first to include model calculations. Figure 1 shows the well-known temperature effect during the first month of the TTAPS 'baseline' simulation of a 5,000-MT nuclear exchange. The maximum temperature decrease of 35°C in this case occurs three to four weeks after the nuclear attack and should be viewed as representative of a value averaged over a hemisphere for a completely land-covered planet. Similarly, TTAPS used a '100-MT cities' scenario in which peak optical depth was about half of their baseline case. Nevertheless, maximum cooling was nearly identical to the baseline case because the absorption of sunlight was effectively saturated. TTAPS¹⁰ suggested that, for the Earth, the large heat capacity of the oceans, which was necessarily neglected in their one-dimensional model, would probably reduce the magnitude of their temperature change curves by about a factor of two, with some variation depending on the proximity of land areas to the ocean. These classical results have been extensively discussed elsewhere^{13,19}.

The second generation of climate models comprised three-dimensional GCMs in which a fixed latitude-height distribution of smoke was assumed. Transport and removal were either ignored or treated by assumption of a given rate of removal independent of the perturbed climate. These models were first used to examine the effect of oceans in moderating the land surface temperature response. Our research group found³⁹, for example, that about 10 to 20 days after imposing smoke with

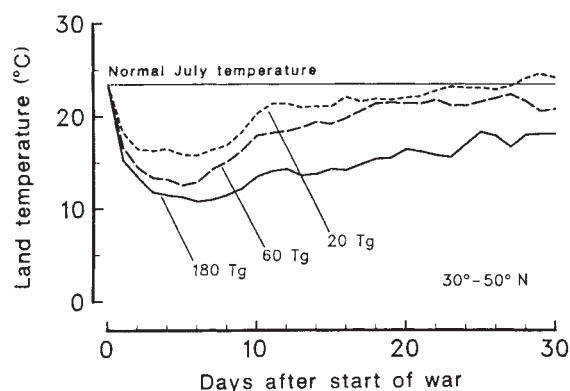


Fig. 2 Mid-latitude average land surface temperatures for July GCM simulations⁹ with three initial smoke amounts as indicated.

visible absorption optical depth of 3 (equivalent to 180 Tg of smoke) over the latitude band 30–70° N, the land surface temperature decrease over this area averaged 25 °C for a July case, ~15 °C for April and <5 °C for January. The April case, which is comparable to the effect of annually averaged solar forcing in one-dimensional models, suggested that the moderating influence of the ocean was indeed about a factor of two. The reason April is comparable to an annually averaged case is that April amounts of solar radiation at each northern latitude are close to mean annual radiation amounts.

Several other important results were obtained from the second-generation models that suggested the need for simulations with more physically realistic models. These include: the severe distortion of atmospheric circulation, suggesting that in spring or summer cases the smoke could be transported into the stratosphere and carried to southern latitudes⁴⁰; probable changes in global precipitation patterns, especially monsoonal rains^{12,20,41}; and a very non-uniform land and sea surface temperature response even under a uniformly thick smoke cloud. Our group⁴² argued that rapid land surface temperature changes were possible, with the potential for sub-freezing land surface temperatures after only a few days of obscured sunlight. Thus, regionally severe effects could occur even with a patchy smoke distribution. It was also discovered³⁹ that air temperatures a few metres above the ground could be several degrees warmer than those of the land surface themselves, implying the need for greatly improved simulations of near-surface atmospheric processes and heat transfer in soil, and the inclusion of a full diurnal cycle of solar forcing^{4,39}. Such factors can influence temperature changes by as much as ± 5 °C depending on season, the assumed smoke distribution and surface properties^{12,43,44}. Finally, it was recognized that radiation models must account for smoke's capability to scatter sunlight and its opacity to the thermal infrared.

Over the past three years all these research topics and model deficiencies have been addressed by one or more of the GCM groups studying the problem, including the UK Meteorological Office, the Computing Centre of the Academy of Sciences (USSR), Los Alamos and Lawrence Livermore National Laboratories (USA), and the National Center for Atmospheric Research (NCAR, USA). Current third-generation climate models include improved radiative transfer calculations compared to earlier GCMs and allow for the interactive transport and removal of smoke by smoke-perturbed circulation systems and by precipitation. Including a process explicitly never guarantees that the quantitative results are improved, but through intercomparison of various models and comparison of results with observations of annual cycle variations, glacial-interglacial transitions and other climate changes we can build confidence in some of the results.

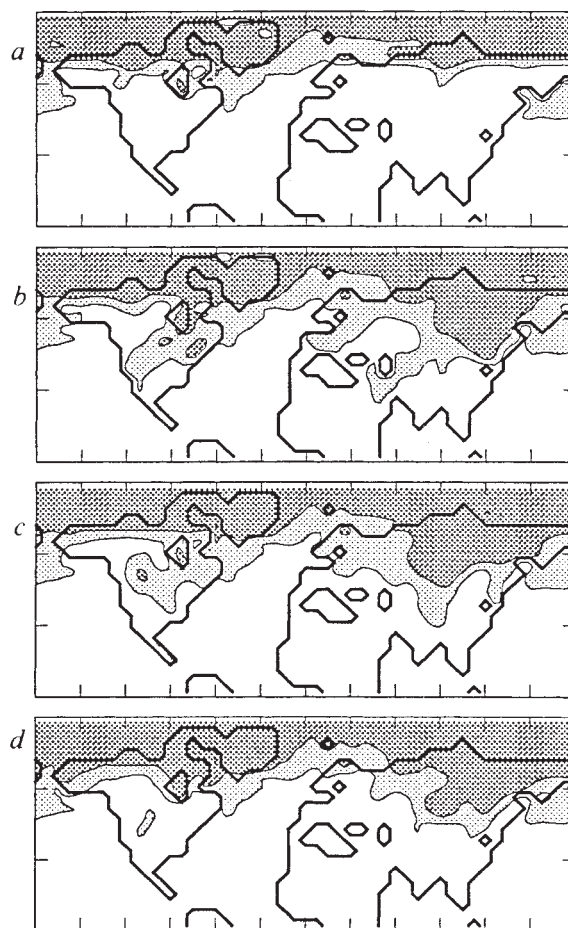


Fig. 3 Land surface temperatures from a July GCM simulation in which 180 Tg of smoke is injected over NATO and Warsaw Pact areas over a period of two days. Dark shading indicates temperatures below 0 °C; lighter shading indicates temperatures between 0 °C and 10 °C. a, Day 0 (initial temperature), b, day 5, c, day 10; d, day 20.

Acute effects

Most of the simulations of the climate effects of nuclear war have focused on the short-term acute changes in the first month or so following the injection of smoke into the atmosphere. Figure 2 shows the NCAR results⁹ for the sensitivity of mid-latitude average land surface temperatures to three smoke injection amounts: 20 Tg, 60 Tg and 180 Tg. The 180-Tg baseline case is equivalent to an injection of 50 Tg of carbon soot; the smaller cases scale linearly. The baseline July case has a maximum mid-latitude, average land temperature drop of ~10 °C below normal, about a factor of two less cooling than found in our second-generation immobile smoke simulation⁴². This is primarily a result of two additional processes: infrared opacity of the smoke providing some 'greenhouse effect' compensatory warming and patchy distribution of smoke, allowing some solar heating in relatively clear periods. Although third-generation refinements have combined to reduce substantially the average land surface temperature perturbations in the northern hemisphere compared to the first-generation of one-dimensional models, these more physically realistic models still produce the rapid cooling episodes found in the earliest GCM results. These results, however, still omit fog formation, the diurnal cycle, soil heat capacity and soil moisture feedbacks. The net effects of these processes involve some degree of compensation, but whether these effects will change regional-scale estimates of acute surface temperature drops by more than a few degrees is

doubtful. Figure 3 shows both the patchy nature of surface temperature response to patchy smoke clouds and the possibility that substantial surface cooling can occur sporadically in the nuclear war zone—even to below freezing in July under appropriate meteorological conditions. Other recent three-dimensional model results also reflect these conditions.

Recently, researchers at the Lawrence Livermore National Laboratory have begun numerical experiments with a 150-Tg soot scenario with smoke of mean absorption optical depth 3 averaged over a hemisphere (M. C. MacCracken and S. Ghan, personal communication). Their preliminary findings indicate a July acute-phase, mid-latitude land surface cooling of $\sim 20^\circ\text{C}$, a result substantially colder than those of 50-Tg soot cases. But this result is still less than the factor of three greater cooling that would be projected if surface temperature depressions were linearly related to optical depth.

Several assumptions in the NCAR calculations⁹ have been challenged^{13,15}, particularly our assumed vertical profile of initial smoke concentration. We assumed that smoke would be injected from the surface to an altitude of seven kilometres with a constant mass mixing ratio (CMR). Our smoke removal scheme (S. L. Thompson and F. Giorgi, manuscript in preparation) was also alleged to be too efficient, removing too much smoke from the atmosphere in the first few weeks of the simulations. Therefore, several additional sensitivity tests were performed, including simulations in which (1) no smoke was removed and (2) smoke was injected from 0–10 km with a constant density concentration profile (CD), thus greatly increasing the average height of initial smoke. Table 1 summarizes the temperature results from these additional simulations¹¹. Even under the assumption of no smoke removal and for the high-altitude CD case, average acute mid-latitude land surface temperature drops of no more than 50% above our baseline values could be obtained, although a greater percentage of the hemisphere would be under the transported smoke. In no case can we produce a mid-latitude mean land surface temperature drop as large as those found in the second-generation studies of 1983–84^{40,42}. It is also important to point out that a 15°C mid-latitude land cooling in an annually averaged model takes temperature to -5°C , whereas an equal cooling in summer in a three-dimensional model will take the temperature to 8°C . That is because the mean annual surface air temperature is about 10°C whereas the July surface temperature is some 13°C warmer. However, the lesser cooling found in three-dimensional models in other seasons will still result in sub-freezing temperatures, at least in mid-latitudes.

One area where there is a large sensitivity to changes in the altitude of smoke injection, is in the removal rate of smoke in the first few weeks of the simulations. In the CMR baseline case about 29% of the smoke remains in the atmosphere after 15 days. In the higher-altitude CD case, however, 50% of the smoke

remains after 15 days. Either result is comparable to the results of other models in which interactive smoke transport, radiative transfer and removal have been included, if they make similar initial smoke assumptions^{12,29}. The sensitivity of the lifetime of smoke in the atmosphere to the smoke injection profile is caused by the interaction between the smoke mass and precipitation, precipitation being the primary means of removal. Smoke injected above the bulk of atmospheric water vapour (above 2–3 km) has a lower probability of removal by rainfall and thus an increased lifetime. The principal effect of increased lifetime of smoke due to the high-altitude CD smoke injection can be seen in Table 1 as increased cooling in the sub-tropics relative to the CMR case.

Smoke that is not removed in the first few days can participate in the self-lofting process whereby absorbed solar energy heats the smoke, induces lifting and drives the smoke higher in the atmosphere. This high-altitude smoke, lofted in spring or summer simulations when available solar energy is at a maximum, forms a very stable heated region where moisture condensation and precipitation are virtually eliminated. In effect, the same process that forms the present stably stratified, heated stratosphere, that is, absorption of sunlight by ozone, works with smoke to create a stable 'smokeosphere'. For typical baseline scenarios this stable layer forms in the region of the current upper troposphere and lower stratosphere, with a bottom boundary at ~ 5 -km altitude²⁹.

The stabilization of the atmosphere by the smoke induces substantial reductions in simulated precipitation. This is demonstrated in Fig. 4, which displays the zonally averaged precipitation rate from the NCAR baseline case. In the first 30 days of the simulation, average rainfall reductions of 20–50% occur in the northern hemisphere mid-latitudes. Even larger percentage reductions can occur over land areas because land areas are more strongly stabilized than ocean owing to low-altitude cooling¹². Such precipitation reductions act as a positive feedback on smoke amount by reducing the rate at which smoke is removed from the atmosphere.

Chronic effects

Long-term (>one month) chronic atmospheric effects would arise from the remnant smoke left in the atmosphere after precipitation removed most of the smoke from the lower to middle troposphere. This remnant smoke could range in amounts from virtually nothing for small smoke injections in winter wars, to 50% or more of the injected smoke from large summer wars involving many urban targets. Thus, the chronic effects of smoke would be quite sensitive to the war scenario, season and smoke injection profile. Acute temperature perturbations have been emphasized over chronic effects because (1) they undoubtedly would be larger in magnitude than long-term effects, and (2) they are easier to simulate because credible long-term calculations require, at least, a model for simulating sea surface temperature changes, sea-ice formation, and high-altitude smoke removal processes. Some preliminary studies^{45,46} have implied that long-term climate changes would be expected given plausible amounts of remnant smoke, but more credible quantitative conclusions must await the next generation of coupled atmosphere–ocean GCMs to be applied to the problem of nuclear smoke. Preliminary calculations with a two-layer GCM at Lawrence Livermore National Laboratory (S. Ghan and M. C. MacCracken, personal communication) do suggest some potentially significant chronic effects, as do some three-dimensional simulations at CSIRO in Australia (B. Pittock, personal communication) and at NCAR.

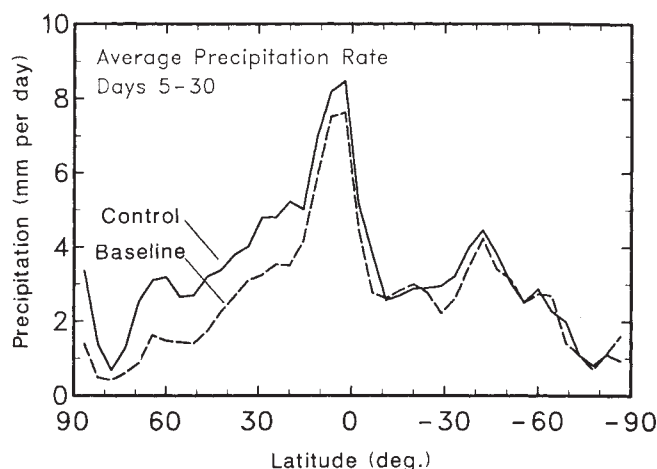
Chronic climate changes would be created by the high-altitude absorption of perhaps 5–40% of incoming solar energy for at least a year. Such perturbations would produce reductions in solar heating of the surface which, in turn, could enhance the probability of late spring or early fall frosts⁹, or could imply strongly reduced Asian or African monsoonal rains through a

Table 1 July land surface temperature difference from control (perturbed minus control) for two vertical injection profiles of smoke and two assumptions of smoke removal

Simulation description		Temperature change ($^\circ\text{C}$)		
Smoke injection profile	Smoke removal	50–70° N	30–50° N	10–30° N
0–7 km CMR	Yes	–12.9	–10.0	–3.9
0–7 km CMR	No	–13.9	–12.9	–5.9
0–10 km CD	Yes	–14.2	–13.2	–8.5
0–10 km CD	No	–14.0	–15.3	–10.6

CMR refers to injection at a constant mass mixing ratio; CD refers to injection at a constant density concentration. Results are averaged over days 5–15, that is, the early 'acute' phase of the post-war effects. Initial smoke loading is 180 Tg, or ~ 50 Tg of carbon soot. (This loading gives an effective hemispheric visible extinction optical depth of 3.88 and absorption optical depth of 1.55.)

Fig. 4 Zonally averaged precipitation rate averaged over days 5–30 of a control July simulation and a baseline 180-Tg smoke injection simulation. The decrease in precipitation in the northern hemisphere of the baseline case is due to heating and stabilization of the atmosphere by smoke.



reduction of the land-sea temperature contrast in summer months²⁰. The potential for a substantial impact on monsoonal rainfall from seemingly modest (5–10%) changes in summer solar radiation reaching the surface has already been demonstrated in the context of palaeoclimatic reconstructions and simulations⁴⁷. Therefore, much of the world's population that would survive the initial effects of nuclear war and initial acute climate effects would have to confront potential agricultural disaster from long-term climate perturbations^{30,31,41}.

The formation of the elevated stable smoke layer may also affect other large-scale environmental threats from nuclear war. Initial indications that heterogeneous oxidation reactions between soot particles and ozone could act as an efficient removal mechanism for stratospheric smoke⁴⁸ have recently been resolved in favour of a relatively long lifetime (of the order of a year) of stratospheric smoke against removal by ozone (S. L. Stephens, unpublished abstract, Defense Nuclear Agency Global Effects Program Technical Review, 1988). It is likely that simulations accounting for the effects of smoke-distorted atmospheric circulations and temperatures will substantially change the current assessments^{1,19,20} of the depletion of stratospheric ozone by nitrogen oxides produced from nuclear fireballs. For example, the self-lofting effect of smoke could drive nitrogen oxides higher in the atmosphere than has been assumed previously. Moreover, the strong heating effect of smoke in the stratosphere could alter temperature-sensitive chemical reaction rates and deplete ozone substantially (S. L. Thompson and P. J. Crutzen, manuscript in preparation). Likewise, intermediate and long-term radioactive fallout rates are dependent on atmospheric stability, precipitation and stratospheric transport. A stabilized atmosphere experiencing large stratospheric circulation perturbations would undoubtedly affect non-local fallout rates, generally in the direction of reduced rates in northern hemisphere mid-latitudes but

increased long-term fallout in the southern hemisphere. Some work has already been done to quantify this effect^{20,49}. One should note, however, that local fallout, which is removed from the atmosphere within a day by gravitational sedimentation, is the most serious fallout problem and would be largely unaffected by smoke perturbations.

Model validity

The validity of current model results requires verification by indirect means because, of course, the purpose of this modelling is to gain insight into the potential effects of nuclear war without direct experimentation. There are several ways of verifying the responses of climate models to large-scale radiative forcing, none of which is entirely satisfactory alone, but when taken together can create a measure of confidence in a model's predictive abilities.

Once a model can simulate the unperturbed climate to some accepted accuracy, the most common means of increasing confidence in model results is to try to eliminate suspected model deficiencies. This is usually done by improving individual model processes and approximations either by comparison with detailed process models of known validity or by comparison with empirical observations of the process in question. This sequential, piecewise approach has been prominent in the nuclear winter problem as shown in Table 2, which gives a summary of significant model improvements and their effects on acute mid-latitude temperature changes. It can be seen that the principal differences between second- and third-generation model results can be attributed to relative warming from (1) smoke patchiness which allows sunlight penetration in clear areas when smoke is selectively removed and transported with the wind instead of being held fixed, and (2) the infrared opacity of smoke which creates a partially mitigating 'greenhouse' effect. (The latter was already included in one-dimensional models.)

Further model improvements are continuing. For example, the NCAR model treatment of the boundary region between the atmosphere and surface ignored the heat capacity of land surfaces and the effects of a stabilized atmosphere on reducing downward transport of heat to the surface³⁹. But three groups have recently included these improvements and found that the errors from the omission of these neglected factors approximately cancel each other^{12,44} (see also G. A. Glatzmaier and R. C. Malone, unpublished abstract, Defense Nuclear Agency Global Effects Program Technical Review, 1987). Moreover, with a diurnal cycle of sunlight included, it was found that the diurnal cycle of surface temperature was severely attenuated for areas that experienced large smoke optical depths^{12,44}. These results suggest that calculated diurnal mean temperatures can be used as first approximations of diurnal temperature minima underneath thick smoke, but that estimates of perturbed diurnal minima should not be made by subtracting calculated diurnal

Table 2 Outline of successive model improvements and their associated effect on acute, mid-latitude (30–50° N) average land surface temperature changes assuming smoke injections equivalent to about 50 Tg of soot carbon

–35 °C	One-dimensional model, annual mean (ref. 10)
	three-dimensional model, July, with:
–17 °C	zonal fixed smoke (ref. 39)
–19 °C	smoke transport (ref. 27)
–13 °C	smoke infrared opacity (ref. 11)
–10 °C	smoke removal (ref. 11)
–13 °C	improved boundary layer (ref. 44)
–10 °C	soil heat capacity (ref. 44)
–8 °C	patchier smoke transport (ref. 12)
(?)	fog formation over land

Values are derived from sources as given.

mean temperature changes from unperturbed diurnal minima.

Some potentially important processes have yet to be included in model simulations of the nuclear smoke problem. A prime example is the formation of fog in the chilled air over land. In general, fog formation would be expected to delay the onset of extreme cold temperatures by keeping near-surface temperatures at the dew point⁵⁰. Fog condensation would release latent heat and effectively spread the intense radiative cooling of the darkened surface over the depth of the fog layer. This would probably tend to ameliorate initial cooling but delay subsequent recovery. It is likely that dew points over most mid-latitude areas in summer would be well above freezing, and thus smoke clouds would have to persist long enough for excess near-surface moisture to be precipitated before sub-freezing temperatures could occur.

A second validating technique is the comparison of results from different models. Although this technique will not uncover fundamental physical errors if they are duplicated in each model, it is valuable for testing the sensitivity of results to model approximations and for finding algorithmic or coding errors. Although it is difficult to compare the complicated GCMs comprehensively, we have compared the temperature results of a simplified smoke scenario as simulated on the NCAR model and a model from Lawrence Livermore National Laboratory¹¹. The differences between the models in acute mid-latitude land surface temperature effects averaged $\pm 9\%$ —about as small as could be expected given the different model resolutions and initial conditions.

Validation by comparison with real climate change is very desirable, but is limited by the dearth of well-documented evidence of causes and effects. Problems with natural analogues to nuclear winter have already been discussed. However, two well understood surrogate climate changes can be used to test model responses to large-scale radiative forcings of comparable or larger magnitude than estimated for nuclear smoke: the observed annual cycle of the seasons and the diurnal cycle of surface climate. The primary cause of cooling from smoke—loss of sunlight at the surface—is also the cause of cooling in winter and at night. Climate models that successfully simulate these changes are more likely to approximate successfully the surface cooling from smoke clouds than models that cannot simulate these natural changes very well. Because of the many simplifications in our and others' climatic models, such testing and verification exercises are always an important part of modelers' research efforts.

Biological effects

It can be misleading to interpret time or spatial averages, as in Fig. 2, without taking into account geographical and weather variability as well. Indeed, only a few hours of having temperatures below some critical level would be enough to affect certain biological processes—for example, sub-freezing temperatures for wheat, or temperatures of 15°C during the flowering phase of rice³⁰. The importance of different temperature thresholds for different biological species at different times emphasizes that analyses of climatological variability need to be better suited to biological applications than standard types of averaging. Such analyses when applied to models could, moreover, be used as an additional tool for the validation of models against observations.

Analyses of variability and extremes are needed not only for the assessment of biological consequences of smoke from a nuclear war, but would be equally useful for assessments of the environmental impacts of increasing 'greenhouse' trace gases^{51–54}. As environmental variability is often more important in biological assessment than mean values, and because the accurate simulation of climate variability depends on the physical realism of models, we believe that it is crucial to involve 'end users' of model results in multi-disciplinary planning of climate simulations of environmental forcings. If, on the other

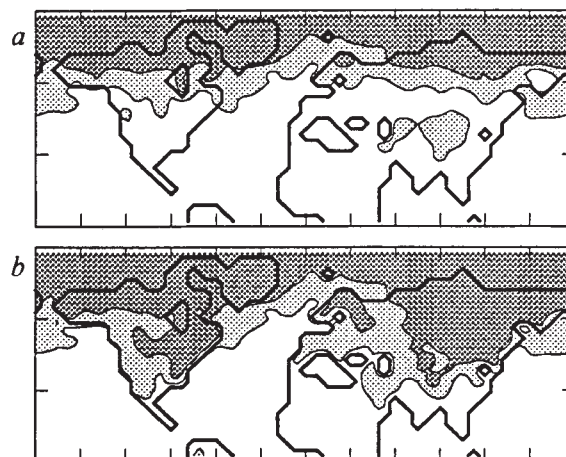


Fig. 5 The coldest land surface temperatures reached at any time during 30-day July GCM simulations. Dark shading indicates temperatures below 0°C ; lighter shading indicates temperatures between 0°C and 10°C . *a*, Unperturbed control case; *b*, perturbed case in which 180 Tg of smoke is injected over NATO and Warsaw Pact areas over the first two days.

hand, assessment is made from physical scientists, then by biologists and finally by social scientists, then the physical scientists could inadvertently invest most of their efforts studying factors of less interest to the other disciplines.

One way to diagnose temperature extremes that are biologically significant is to display the coldest temperature reached at any time during a simulation. This is done for a 30-day July control case and the smoke-perturbed CMR baseline case in

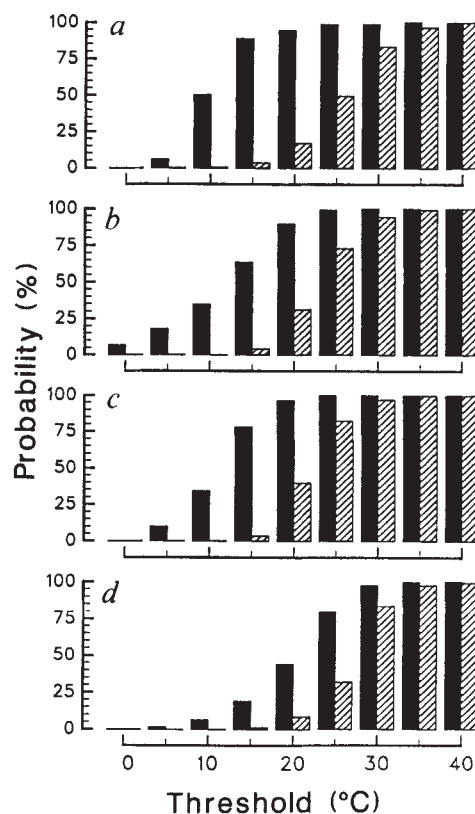


Fig. 6 Cumulative probability distributions for regional surface temperatures falling below threshold values in July GCM simulations. Hatched bars: unperturbed control case. Solid bars: 180-Tg smoke-perturbed CD case. *a*, US midwest region; *b*, USSR Ukraine region; *c*, western European region; *d*, southeastern China region.

Fig. 5. As our model runs do not yet include a diurnal cycle, these are daily average representations of extreme variability. Note that sub-freezing extremes normally confined to polar latitudes or high mountain areas in the control simulation have expanded into mid-latitudes in the baseline simulation, implying a substantial probability of crop-damage by frosts in regions normally frost-free in the middle of the growing season.

The study of climate effects on water supplies⁵⁵ or agriculture^{51,52} clearly shows that a change in the probability of extreme events is of great significance. Figure 6 shows histograms of the cumulative probability of temperatures⁵⁶ falling below threshold values for selected geographic regions. In a simulation of usual July conditions in the mid-western United States (Fig. 6a), there is a 100% probability of surface temperatures falling below 40 °C but 0% probability that diurnally averaged temperatures will fall below 15 °C. The smoke-perturbed case, in contrast, shows a substantial increase in the probability of cooler temperatures, although no sub-freezing samples are found. Of course, because our GCM is in fact a weather model with substantial weather fluctuations⁵⁷ there is still considerable sampling uncertainty in these results, which are given primarily to display the methods and are not to be taken quantitatively at face value as climate statistics. This caveat is underscored by results from the baseline CMR case (not shown) where 3% of the sampled surface temperatures in the US mid-west are below 0 °C although the mean latitudinal surface cooling is less for this case (see Table 1). For the Ukraine region (Fig. 6b), sub-freezing samples occur about 7% of the time in the particular CD case. In the analogous CMR case (not shown) there are no sub-freezing samples. Histograms are also plotted for western Europe and south-east China, the latter principally because it shows a 20% chance of surface temperatures below the agriculturally significant 15 °C threshold³⁰.

Conclusions

The original estimates of severe, long-term, widespread northern hemisphere temperature declines following a large nuclear war have been mitigated by successive generations of more comprehensive climate models. Maximum summertime, northern hemisphere, average land surface temperature changes of 5–15 °C for typical baseline smoke injection scenarios—similar to the mid-latitude change from summer to autumn—have replaced the winter-like estimates of 25–30 °C surface cooling over land.

Sporadic cooling events are still simulated by all modelling groups for mid-latitude land areas. Because autumn is the time that the growing season ends in the mid-latitudes of the northern hemisphere and the summer monsoons fade in Asia and northern Africa, we are willing to continue to use our 'nuclear fall' metaphor as an appropriate shorthand label for the magnitude and biological implications of the results from present models given presently accepted baseline estimates by smoke production. Of course, neither 'nuclear winter' nor 'nuclear fall' are precise descriptions of the range of climate statistics associated with these seasons.

Current climate model simulations for baseline cases suggest that it is unlikely for the climate effects of nuclear war alone to be more devastating to combatant nations than the direct effects of the use of many thousands of nuclear weapons. However, on a global scale, all the indirect environmental effects of a large nuclear war would, to some degree, act synergistically and interact with sociological effects, such as the disruption of international trade in energy, food, fertilizer, medicine and spare parts^{9,30,58,59}. For example, the interaction between a failure of the normal Asian summer monsoon together with the virtual elimination of international food trade would be a prescription for mass starvation on the Indian subcontinent. Many other interactions such as the effects of dramatic stratospheric ozone reductions, are subjects for speculation but such synergisms deserve serious study to allow us to draw less tentative conclusions. Therefore, it remains plausible that the sum of climate disturbances, radioactive fallout, ozone depletions and the interruption of basic social services, could threaten more people globally than would the direct effects of nuclear explosions. The latter is a sufficiently horrendous prospect to make a nuclear war an unprecedented human catastrophe. The addition of environmental effects deepens the horror for the combatants and significantly extends the catastrophic consequences to non-combatants.

We thank C. Covey, F. Giorgi and V. Ramaswamy for their collaboration. We acknowledge useful suggestions and criticisms from referees Michael MacCracken, Carl Sagan and Richard Turco. Our work was partially sponsored by the Defense Nuclear Agency. NCAR is sponsored by the NSF. Opinions, findings, conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the NSF or Defense Nuclear Agency.

- Crutzen, P. J. & Birks, J. W. *Ambio* **11**, 115–125 (1982).
- Teller, E. *Nature* **310**, 621–624 (1984).
- Singer, S. F. *Nature* **310**, 625 (1984).
- Thompson, S. L., Schneider, S. H. & Covey, C. *Nature* **310**, 625–626 (1984).
- Sagan, C. *Nature* **317**, 485–488 (1985).
- Sagan, C. *Foreign Affairs* **62**, 257–292 (1983).
- Wohlstetter, A. *Foreign Affairs* **63**, 962–994 (1985).
- Rathjens, G. W. & Siegel, R. H. *Issues in Science and Technology* 123–128 (Winter 1985).
- Thompson, S. L. & Schneider, S. H. *Foreign Affairs* **64**, 981–1005 (1986).
- Turco, R. P., Toon, O. B., Ackerman, T. P., Pollack, J. B. & Sagan, C. *Science* **222**, 1283–1292 (1983).
- Thompson, S. L., Ramaswamy, V. & Covey, C. *J. geophys. Res.* **92**, 10942–10960 (1987).
- Ghan, S. J., MacCracken, M. C. & Walton, J. J. *J. geophys. Res.* (in press).
- Thompson, S. L. *et al. Foreign Affairs* **65**, 163–178 (1986).
- Nature* **323**, 95–96 (1986).
- Thompson, S. L. *et al. Environment* **28**, 2–5 (1986).
- Warner, F. *et al. Environment* **29**, 4–5 (1987).
- Glieck, J. *The New York Times* 1 (22 June 1986).
- Marshall, E. *Science* **235**, 271–273 (1987).
- U.S. National Research Council *The Effects on the Atmosphere of a Major Nuclear Exchange* (National Academy Press, Washington, 1985).
- Pittcock, A. B. *et al. Environmental Consequences of Nuclear War, Volume I: Physical and Atmospheric Effects* (Wiley, New York, 1986).
- Barnaby, F. *et al. Ambio* **11**, 94–99 (1982).
- Crutzen, P. J., Brühl, C. & Galbally, I. E. *Clim. Change* **6**, 323–364 (1984).
- Small, R. D. & Bush, B. W. *Science* **229**, 465–469 (1985).
- Cotton, W. R. *Am. Sci.* **73**, 275–280 (1985).
- Golding, B. W., Goldsmith, P., Machin, N. A. & Slingo, A. *Nature* **319**, 301–303 (1986).
- Penner, J. E. *Nature* **324**, 222–226 (1986).
- Thompson, S. L. *Nature* **317**, 35–39 (1985).
- Malone, R. C., Auer, L. H., Glatzmaier, G. A., Wood, M. C. & Toon, O. B. *Science* **230**, 317–319 (1985).
- Malone, R. C., Auer, L. H., Glatzmaier, G. A., Wood, M. C. & Toon, O. B. *J. geophys. Res.* **91**, 1039–1053 (1986).
- Harwell, M. A. & Hutchinson, T. C. *Environmental Consequences of Nuclear War, Volume II: Ecological and Agricultural Effects* (Wiley, New York, 1985).
- Ehrlich, P. R. *et al. Science* **222**, 1293–1300 (1983).
- Alvarez, L. W., Alvarez, W., Asaro, F. & Michel, H. V. *Science* **208**, 1095–1108 (1980).
- Pollack, J. B., Toon, O. B., Ackerman, T. P. & McKay, C. P. *Science* **219**, 287–289 (1983).
- Officer, C. B., Hallam, A., Drake, C. L. & Devine, J. D. *Nature* **326**, 143–149 (1987).
- Wolbach, W. S., Lewis, R. S. & Anders, E. *Science* **230**, 167–170 (1985).
- Schneider, S. H. *Clim. Change* **5**, 111–113 (1983).
- Catchpole, A. J. W. & Faurer, M. *Clim. Change* **5**, 114–139 (1983).
- Lough, J. M. & Fritts, H. C. *Clim. Change* **10**, 219–239 (1987).
- Covey, C., Thompson, S. L. & Schneider, S. H. *J. geophys. Res.* **90**, 5615–5628 (1985).
- Thompson, S. L. *et al. Ambio* **13**, 236–243 (1984).
- Schneider, S. H. & Londer, R. *The Coevolution of Climate and Life* 359 (Sierra Club Books, San Francisco, 1984).
- Covey, C., Schneider, S. H. & Thompson, S. L. *Nature* **308**, 21–25 (1984).
- Cess, R. D., Potter, G. L., Ghan, S. J. & Gates, W. L. *J. geophys. Res.* **90**, 12937–12950 (1985).
- Mitchell, J. F. B. & Slingo, A. *J. geophys. Res.* (submitted).
- Robock, A. *Nature* **310**, 667–670 (1984).
- Covey, C. *Nature* **325**, 701–703 (1987).
- Kutzbach, J. E. & Street-Perrott, F. A. *Nature* **317**, 130–134 (1985).
- deHaas, N., Fristrom, R. M., Linevsky, J. J. & Silver, D. M. *Johns Hopkins, APL tech. Digest* **7**, 181–186 (1986).
- Knox, J. & Shapiro, C. *Nature* **321**, 21 (1986).
- Molenkamp, C. R. in *Aerosols and Climate* (eds Hobbs, P. B. & McCormick, M. P.) (Deepak Pub. Co., Hampton, in the press).
- Parry, M. L. & Carter, T. R. *Clim. Change* **7**, 95–110 (1985).
- Mearns, L. O., Katz, R. W. & Schneider, S. H. *J. Clim. appl. Meteorol.* **23**, 1601–1613 (1984).
- Schneider, S. H. & Thompson, S. L. in *Report of the 1984 Conference on the Global Possible* (ed. Repetto, R.) 397–430 (World Resources Institute, Washington, 1985).
- Bolin, B., Doos, R., Jager, J. & Warrick, R. A. (eds) *The Greenhouse Effect, Climatic Change and Ecosystems SCOPE 29* (Wiley, Chichester, 1986).
- Schwarz, H. E. in *Climate, Climatic Change and Water Supply* 111–129 (National Academy Press, Washington, 1977).
- Reed, D. N. *J. Climatol.* **6**, 233–253 (1986).
- Chervin, R. M. *J. Atmos. Sci.* **43**, 233–251 (1986).
- Ehrlich, P. R. & Ehrlich, A. H. *Population, Resources, Environment* 2nd edn (Freeman, San Francisco, 1972).
- Schneider, S. H. & Mesirov, L. E. *The Genesis Strategy: Climate and Global Survival* 204–205 (Plenum, New York, 1976).

Axi-symmetric Earth models and inner-core anisotropy

Peter M. Shearer*, Kenneth M. Toy† & John A. Orcutt†

* Department of Earth Sciences, University of Cambridge, Bullard Laboratories, Madingley Road, Cambridge CB3 0EZ, UK

† Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093, USA

Seismic waves passing through the middle of the Earth travel slightly faster in a N-S direction than in an E-W direction. A comparison of travel times for different P-wave phases shows that most of this anomaly is within the inner core, and that uniform anisotropy of the inner core is the most likely cause.

BODY wave travel times as tabulated by the International Seismological Centre (ISC) are increasingly being used to formulate three-dimensional models of mantle^{1,2}, core-mantle boundary^{3,4}, and inner-core structure^{5,6}. Despite sophisticated inversion techniques, these studies produce models that often do not agree, even at very low-order terms, in their spherical harmonic expansions. It is difficult to evaluate the relative accuracy of the different models, because they result from relatively complicated processing techniques, which typically involve both corrections (that is, ellipticity, station, upper- and lower-mantle structure, core-mantle boundary structure) and smoothing (namely, averaging residuals within cells, spherical harmonic expansions, stochastic filtering).

One of the most serious unresolved questions in these analyses concerns the origin of the large C_{20} (spherical harmonic coefficient) anomaly. It is clear from PKIKP travel-time studies^{5,6} and normal-mode splitting studies⁷⁻⁹ that, on average, the Earth is slightly faster in a N-S direction than in an E-W direction for rays and modes that sample deep within the earth. For near-vertical PKIKP rays, this difference is about 2 s. This suggests the presence of an axi-symmetric C_{20} term in the pattern of lateral heterogeneity in the Earth, but the exact depth of this anomaly is controversial. Ritzwoller *et al.*⁷ place the anomaly in the lower mantle and outer core, Creager and Jordan³ place it at the core-mantle boundary (CMB), whereas Morelli *et al.*⁶ argue that it must be within the inner core and go on to suggest that inner-core anisotropy, rather than lateral heterogeneity, may be responsible. For all of these models, the C_{20} term is much larger than any of the other aspherical terms. Thus the question of the depth of the C_{20} anomaly must be resolved before comparisons of higher-order terms in the models will be meaningful.

Our analysis will focus on the key question of the depth of the C_{20} anomaly. Accordingly, we will consider only axi-symmetric Earth models, examining patterns of travel-time residuals averaged over longitude, and then comparing these patterns with those predicted by the various models. We will generally attempt to remove as much as possible of the bias introduced by the uneven distributions of sources and receivers, and smooth the data in order to concentrate only on the low-order terms. Complicated inversion schemes will be minimized, with corrections applied to the data only if they significantly affect the results.

We will conclude that the data strongly indicate a C_{20} anomaly within the inner core, and that inner-core anisotropy, rather than lateral heterogeneity, is the most probable cause. Our preferred anisotropy model is hexagonally symmetric with a fast N-S symmetry axis, and is approximately uniform throughout the inner core.

Procedure

A comparison of PKP(BC) and PKIKP residuals provides the

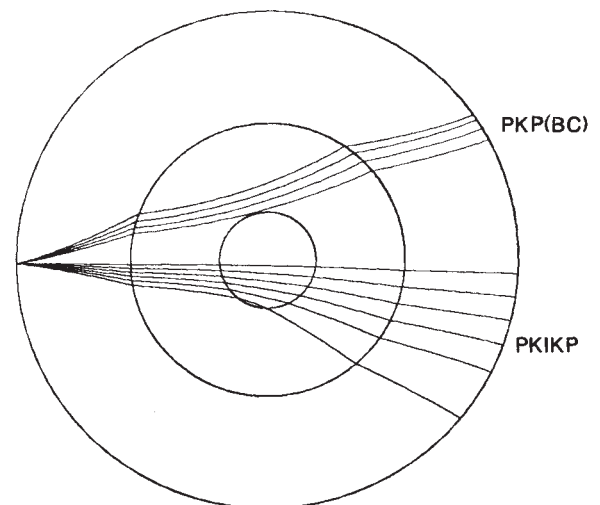


Fig. 1 Ray paths for PKP(BC) and PKIKP phases.

most direct method of resolving the question of whether the C_{20} anomaly exists within or outside the inner core. If the anomaly is within the inner core, then the PKIKP residuals should show a C_{20} pattern, whereas the PKP residuals would not show this pattern. If the anomaly is outside the inner core (for example, at the core-mantle boundary), then both phases should show an anomalous pattern. The BC branch of PKP is most applicable for this comparison because it turns within the lowest part of the outer core. Figure 1 plots ray paths for PKP(BC) and PKIKP, and illustrates that the ray paths through the mantle and outer core are roughly the same for the two phases.

Figure 2a shows PKP(BC) travel-time residuals as a function of range for 174,028 PKP travel times (ISC 1964 to 1984). Residuals are calculated relative to the PR Earth model¹⁰ and corrected for ellipticity. No station corrections or mantle model corrections are applied. The PR Earth model, as noted by Toy *et al.*¹¹, produces residuals with a mean of several seconds for deep body-wave phases. However, we are not concerned with radial Earth model deviations in this analysis, and so will only consider deviations from the mean residual. We attempt to minimize contamination from the AB and EF branches of PKP by using the relatively narrow four-second data window shown by the white lines in Fig. 2a. The turning-point radius for the BC branch of PKP varies from 1,221.5 km (the inner-core radius) to about 1,900 km. Figure 2b is a similar plot of 257,429 PKIKP travel-time residuals. As noted by Poupinet *et al.*⁵, the arrivals are contaminated at ranges between 140° and 154° by the AB and BC branches and possible diffracted energy from the B caustic. We window the data as shown to remove this contamination; the missing data correspond to turning point radii of 905 to 1,155 km.

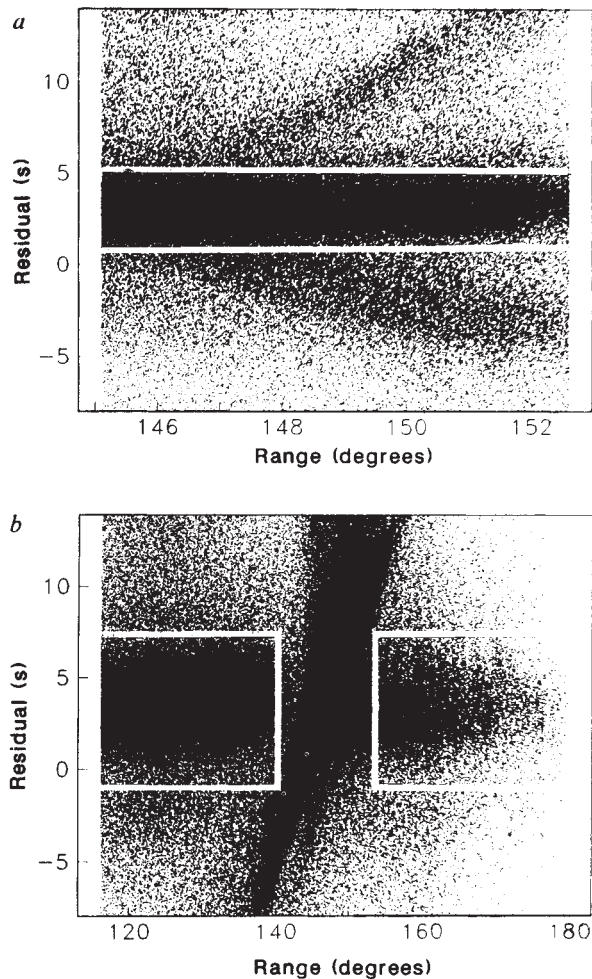


Fig. 2 Travel-time residuals relative to the PR Earth model for: *a*, PKP(BC) arrivals, windowed as shown between 1 and 5 s in order to isolate the BC branch from the AB branch above and the EF branch below; *b*, PKIKP arrivals windowed as shown between -0.8 s and 7.2 s, excluding data between 140 and 150° which are contaminated by the AB and BC branches of PKP.

To show the axi-symmetric signal as a function of depth for these data, we plot the data in two different ways. First, we plot travel-time residuals versus the colatitude of the ray turning point. This type of plot is most sensitive to lateral heterogeneity at the turning point depth. Second, we plot the residuals versus the ray angle from the N-S axis at the turning point. This type of plot is most sensitive to hexagonally symmetric anisotropy of the type proposed for the inner core by Morelli *et al.*⁶. Changes in the shape of these residual curves with depth provide a direct method of resolving between inner-core anisotropy and lateral heterogeneity.

Because of the very uneven distribution of sources and receivers for the ISC data set, many studies have attempted to even out the data as much as possible by weighting observations from 'rare' ray paths proportionately more than data from 'common' ray paths. These averaging schemes typically take the form of dividing the data into bins which depend upon various parameters (such as source, receiver, or turning point location), and then averaging residuals within each bin, thus forming a reduced data set with a more uniform distribution than the complete data set. Assuming a reference radial Earth model and a given seismic phase, there are four independent parameters that completely describe the ray path. For example, the source and receiver locations (latitude and longitude) would specify the ray path. For our analysis, the turning-point depth and location, and the local ray angle at the turning point are the

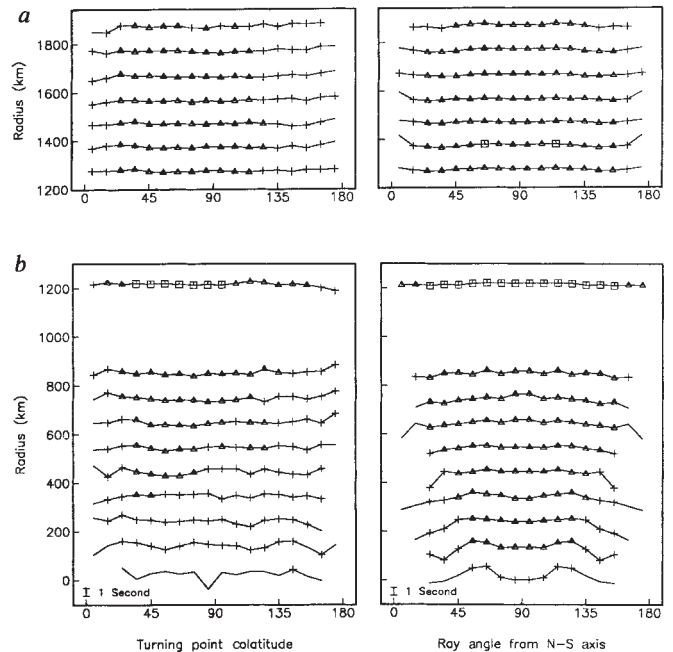


Fig. 3 Axi-symmetric residuals for *a*, PKP(BC) and *b*, PKIKP plotted as a function of turning point colatitude and ray angle from the N-S axis. 100 km on the vertical scale is equivalent to a 3 s residual. Early arrivals are indicated by negative residuals, so, for example, a curve which is concave downward indicates faster arrivals at 0° and 180° than at 90°. The missing data between 900 and 1,150 km radius correspond to the contaminated region shown in Fig. 2*b*. Symbols indicate the number of 'independent' ray paths per data point, with squares indicating more than 200, triangles more than 50 and crosses more than 10. Data points with less than 4 paths are not plotted.

most important, and it is these four parameters that we use to sort and bin the data.

Our averaging scheme has the following form: first we divide the data by turning-point depth in intervals of 100 km. Second, we consider the turning-point latitude and longitude, and sort the data into 416 bins of roughly equal area, with a latitude spacing of 10° and a longitude spacing varying from 10° at the equator to 120° near the poles. Finally, within each bin, we further divide the data into 18 intervals of ray azimuth (measured from the local north on the surface of a reference sphere). Note that in general this ray azimuth differs from the ray angle from the N-S axis which we will eventually plot. The angles are related by $\theta = \cos^{-1}(\sin \phi \cos \omega)$, where θ is the ray angle from the N-S axis, ϕ is the colatitude of the turning point, and ω is the local ray azimuth. Because of the ray path reciprocity between sources and receivers, we consider local turning-point azimuths from 0° to 180° rather than from 0° to 360°. The averaging over longitude causes the ray-angle plots to be symmetric about $\theta = 90^\circ$.

Thus, for each depth interval, we tabulate values for a 416 by 18 matrix, containing the average residual within each bin. This reduced data set forms the basis for the analysis which follows. Within the limitations of the finite number of bins, this sorting scheme is independent of the choice of coordinate axes. Because of the uneven distribution of sources and receivers, the matrix is very sparse, with only about 10% of the bins containing data. To plot travel-time residual versus turning-point colatitude, we simply average data from bins with the same colatitude. For the plots of residual versus ray angle from the N-S axis, we use the above equation to calculate this ray angle and then average all bins within 10° intervals. Each data point plotted in this paper is an average of from 4 to over 200 individual bins (see Fig. 3), where each bin represents a distinctly different ray path.

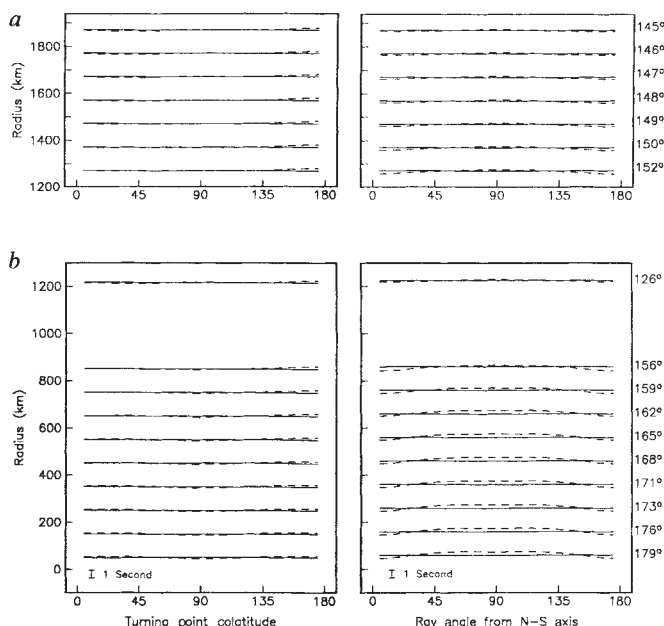


Fig. 4 Axi-symmetric *a*, PKP(BC) and *b*, PKIKP residuals predicted by models of heterogeneity at the core-mantle boundary from Morelli and Dziewonski⁴ (solid line) and Creager and Jordan³ (dashed line), shown at intervals of 100 km in turning-point depth. Approximate ranges corresponding to these turning-point depths are shown at the right. 100 km on the vertical scale is equivalent to a 3 s residual.

Figure 3 shows these residual patterns as a function of turning-point depth. PKIKP data are missing at radii of 900 to 1,155, corresponding to the contaminated region shown in Fig. 2b. The uppermost PKIKP bin contains data at depths between 1,155 and 1,221.5 km (the inner-core radius), with most of the data in the top 10 km. The strongest anomaly pattern is visible in the ray-angle plot of PKIKP, with the polar arrivals fast compared to the equatorial arrivals, with this difference generally increasing with depth in the inner core. In contrast, the PKP(BC) residual plots show very little variation. This is direct evidence that most of the C_{20} travel-time anomaly is probably within the inner core. The lack of a consistent pattern in the PKIKP turning-point colatitude plot indicates that the anomaly pattern is more compatible with inner-core anisotropy than with lateral heterogeneity.

Model comparison

We compare these results with those predicted by several different models of the inner core and core-mantle boundary. These models include:

(1) The Morelli and Dziewonski⁴ model of core-mantle boundary topography, derived from ISC PcP and PKP(BC) phases using a spherical harmonic expansion up to degree 4. Although the CMB varied by up to 6 km in this model, the axi-symmetric terms are very small ($C_{10}=0.34$, $C_{20}=0.14$, $C_{30}=-0.26$, $C_{40}=-0.11$ km).

(2) The Creager and Jordan³ model of CMB heterogeneity derived from PKP(AB) and PKIKP phases. Axi-symmetric terms in the spherical harmonic expansion of their model are $C_{10}=-0.042$, $C_{20}=-0.104$, $C_{30}=0.028$, $C_{40}=-0.048$, $C_{50}=-0.011$ s (Morelli and Dziewonski⁴ normalization). These numbers represent the travel-time perturbation at the CMB for vertically travelling rays through core-side heterogeneity (their preferred model). A 0.1-s difference in vertical travel time at the CMB boundary is equivalent to about 2 km of CMB topography, so clearly this model is very different from the Morelli and Dziewonski model discussed above.

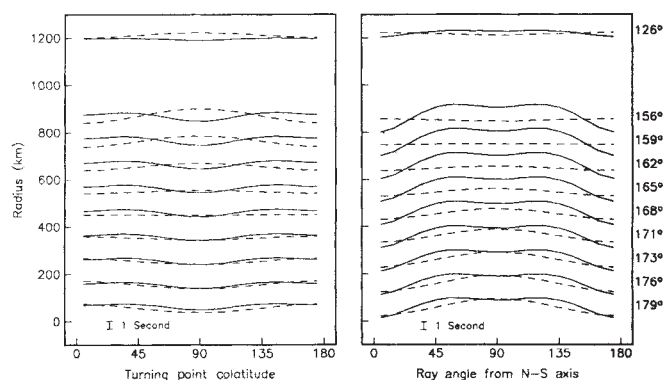


Fig. 5 Axi-symmetric PKIKP residuals predicted by a model of inner-core anisotropy from Morelli *et al.*⁶ (solid line) and an *ad hoc* of inner-core heterogeneity (dashed line). Scaling is as in Figs 3 and 4.

(3) The Morelli *et al.*⁶ model of inner-core anisotropy. In this model, velocities in the inner core vary as

$$V = V_{eq}(1 + \varepsilon \cos^2 \theta + \sigma \cos^2 \theta \sin^2 \theta) \quad (1)$$

where V_{eq} is the equatorial velocity and θ is the ray angle from the N-S symmetry axis. This perturbation tapers with depth as $(r/R_{IC})^2$ in the inner core so that the model is isotropic at the centre of the Earth. Morelli *et al.* inverted near vertical PKIKP data (170–180° range) to obtain $\varepsilon=0.032$ and $\sigma=-0.064$, numbers which imply a 3% difference in polar versus equatorial P -wave velocities near the surface of the inner core.

(4) A simple *ad hoc* model of inner-core heterogeneity. For this model, we assume a velocity perturbation in the inner core defined simply by a C_{20} term which tapers with depth as $(r/R_{IC})^2$. At the surface of the inner core, $C_{20}=0.1$ km s⁻¹, resulting in a model in which the inner core is fast near the poles and slow at the equator.

For vertical PKIKP rays, these models predict polar versus equatorial travel-time differences of -0.03, 0.88, 2.3 and 2.2 s, respectively. We calculate the effect of these models by tracing rays through the PR Earth model¹⁰, rotating representative rays to different turning point positions and azimuths, summing along each path the residuals predicted by the models, and then averaging the residuals within the appropriate bins. The resulting residual patterns as a function of turning-point depth are shown in Figs 4 and 5.

The residuals predicted by the Morelli and Dziewonski⁴ CMB model are barely noticeable at the scale of the plots (Fig. 4). The Creager and Jordan³ CMB model predicts ray-angle residuals that appear to fit the more shallow turning PKIKP data (Fig. 3), but these residuals do not increase with depth to match the large PKIKP residuals deeper within the inner core. The Creager and Jordan model did not consider PKP(BC) data and was derived partially from PKIKP data, assuming inner-core homogeneity. The model predicts a PKIKP residual pattern similar in shape (but smaller by a factor of about 3) to that predicted by the Morelli *et al.*⁶ model of inner-core anisotropy. If the Creager and Jordan model were increased in size to approximate the PKIKP residuals of the Morelli *et al.* model, the resulting PKP(BC) residuals would become too large to be consistent with the data. This indicates that most of the anomaly shown in the PKIKP data must be within the inner core, not at the core-mantle boundary. The relatively low-amplitude PKP(BC) residual pattern (Fig. 3) is roughly compatible with the pattern of both CMB models (Fig. 4). The Creager and Jordan model gives a somewhat better fit to the positive residuals at turning-point colatitudes near 180° and to the slightly concave-downward shape of the ray-angle residuals between 45° and 135°, whereas the Morelli and Dziewonski model gives a better

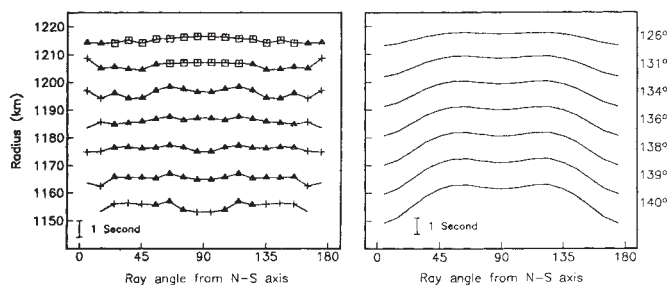


Fig. 6 Ray angle residuals for shallow turning PKIKP rays, shown at intervals of 10 km in depth, compared with predicted results for the Morelli *et al.*⁶ anisotropy model. Note the expanded time scale compared to the earlier plots; 1 s is equivalent to 6 km. Symbols are as in Fig. 3.

fit to the ray-angle data near the poles. But these PKP(BC) plots are most sensitive to anomalies at the turning-point depths (that is, in the lower part of the outer core), and are not ideal for comparing CMB structures. The relatively flat shape of the PKP(BC) ray-angle residual pattern argues against the recent hypothesis of anisotropy in the lower part of the outer core¹².

Figure 5 shows that the Morelli *et al.*⁶ inner-core anisotropy model and our *ad hoc* heterogeneity model predict similar PKIKP residual patterns for the deepest turning rays. However, the predicted patterns diverge for the more shallow turning rays. For PKIKP rays turning at radii from 600 to 800 km, the 2θ pattern in the ray-angle data is more consistent with the anisotropy model than the flat curve predicted by the heterogeneity model. The turning-point colatitude plots also favour the anisotropy model, with the concave-upward residual curves at radii of 600 to 800 km agreeing with the anisotropy model, but disagreeing sharply with the concave-downward curves predicted by the heterogeneity model. In addition, at deeper depths (radii from 100 to 300 km) the anisotropy model

better reproduces the flattening in the residual curve near the equator.

The Morelli *et al.*⁶ model, however, does not accurately predict the lessening of the ray-angle anomaly pattern with increasing radius in the inner core (compare Figs 3 and 5 at intermediate depths). This is because the $(r/R_{IC})^2$ factor in their model concentrates the anomaly near the surface of the inner core. Morelli *et al.*⁶ preferred this model over a uniformly anisotropic inner-core model, because it agrees better with normal-mode splitting data⁸, but our data show that such strong anisotropy near the surface of the inner core is incompatible with the PKIKP residual pattern for shallow turning rays.

This discrepancy can be seen most clearly if we examine the uppermost bin at 10 km intervals. Figure 6 is a close up of the top 70 km of the PKIKP residuals, showing a small but discernible 2θ pattern in the ray-angle plots. Figure 6 also shows the predicted anomaly patterns for the Morelli *et al.* anisotropy model. Despite the shallow turning depths in the inner core, the model predicts very large residual anomalies, a result of the $(r/R_{IC})^2$ taper in the model concentrating the perturbations toward the surface of the inner core. The resulting residual patterns are clearly incompatible with the data.

A uniform anisotropy model

A better fit to the PKIKP residual pattern can be obtained by assuming uniform anisotropy throughout the inner core. To obtain such a model, we consider velocity perturbations of the form

$$V = A + B \cos 2\theta + C \cos 4\theta \quad (2)$$

where θ is the ray angle from the N-S symmetry axis. Note that this is equivalent to equation (1), where $A = V_{eq}(1 + \epsilon/2 + \sigma/8)$, $B = V_{eq}\epsilon/2$, $C = -V_{eq}\sigma/8$. When the anisotropy is weak ($B, C \ll A$), this expression is a good approximation to the equation first derived by Backus¹³ and discussed extensively by Crampin¹⁴: $V^2 = A' + B' \cos 2\theta + C' \cos 4\theta$, where $A' = A^2$, $B' = 2AB$, $C' = 2AC$. Assuming a (001) hexagonal symmetry axis, we also have expressions for some of the elastic constants^{13,14}: $\Gamma_{3333} = A' + B' + C'$, $\Gamma_{1111} = \Gamma_{2222} = A' - B' + C'$, $\Gamma_{1133} + 2\Gamma_{1313} = A' - 3C'$, where Γ_{ijkl} is the density-normalized elastic tensor.

Using equation (2), we perform a simple least-squares fit to the PKIKP ray-angle data shown in Fig. 3, assuming no depth dependence for the anisotropy. The resulting values ($B = 0.054$, $C = 0.024 \text{ km s}^{-1}$) are sufficient to reduce the variance of the ray-angle residuals by 68%. Predicted PKIKP residual curves for our model are compared with the data in Fig. 7a. The ray-angle fit is excellent at most depths, except for some mismatch near the endpoints (0 and 180°) and for the deepest turning rays (0–200 km radius). These are sparse areas of the data set, where we might expect the results to be somewhat noisier.

The predicted colatitude residual curves correspond less well to the data, but are still in reasonable agreement. Particularly encouraging is the reproduction of the slightly concave-upward shape of the residual curves at radii between 600 and 800 km. As a further test of our anisotropy model, we compare predicted residuals versus the data for the shallow turning PKIKP rays (see Fig. 7b). The ray-angle residual agreement is good at all depths, despite the fact that our inversion uses only the uppermost bin (1,210–1,221.5 km).

The agreement between our model and the data should not be exaggerated as the 68% variance reduction applies only to the binned and averaged data, not to the 257,429 individual PKIKP travel-time residuals, which are, of course, much noisier. The ray-angle plots are most sensitive to anisotropy, and exhibit the closest fit to the data. The colatitude plots fit less well, and indicate that some lateral heterogeneity may also be present in the inner core. In general, however, it is clear that the bulk of the axi-symmetric residual pattern can be fit with a uniform anisotropy model involving just two free parameters.

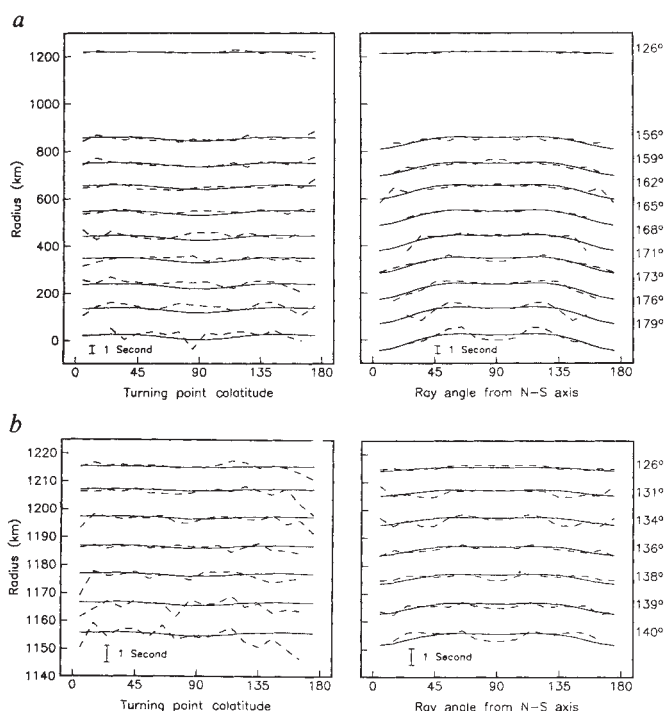


Fig. 7 PKIKP residuals predicted by a simple model of uniform inner-core anisotropy (solid line), compared with the data (dashed line) for the entire inner core, *a*, and for the uppermost 70 km of the inner core, *b*. Scaling is as in Figs 3 and 7.

The relatively large size of the 4θ term in our model (46% of the 2θ term) is what causes the equatorial flattening in the ray-angle residual data curves shown in Fig. 7. The data also consistently exhibit this flattening, suggesting that a significant 4θ term is a real feature of inner-core anisotropy. If inner-core anisotropy is a result of preferred crystal alignment (which seems the most probable cause), then the observed ratio of about 2 between the 2θ and 4θ coefficients may provide an important constraint on the crystal structure of the inner core.

Our anisotropy model ($B = 0.054$, $C = 0.025 \text{ km s}^{-1}$) predicts anisotropic P -wave velocity variations of slightly under 1% at the surface of the inner core. In attempting to reconcile normal-mode splitting data with PKIKP travel-time data, Woodhouse *et al.*⁸ noted that the anisotropy required at the surface of the inner core to explain the mode splitting is about a factor of two larger than the anisotropy indicated by the PKIKP residuals. This comparison assumed the Morelli *et al.*⁶ PKIKP anisotropy

model, in which the anisotropic velocity variations are over 3% near the surface of the inner core. However, we have shown that such large velocity variations at these depths predict residual patterns which are incompatible with the PKIKP data for rays turning at shallow to intermediate depths within the inner core. Our preferred uniform model of inner core anisotropy is ~ 6 times too small to explain the normal-mode splitting data. Although we cannot rule out the possibility that a more complicated model of inner-core anisotropy and/or heterogeneity might still reconcile the PKIKP and normal-mode data sets, it seems unlikely that the primary source for the C_{20} anomaly pattern in the normal-mode data is within the inner core. A key question for future research will be to resolve between the different models^{3,4,7} which have been proposed for the core-mantle boundary to see if a single model can be found for this region which is compatible with both normal-mode and travel-time data.

Received 30 December 1987; accepted 21 March 1988.

1. Clayton, R. W. & Comer, R. P. *Trans. Am. geophys. Un.* **64**, 776 (1983).
2. Dziewonski, A. M. *J. geophys. Res.* **89**, 5929–5952 (1984).
3. Creager, K. C. & Jordan, T. H. *Geophys. Res. Lett.* **13**, 1497–1500 (1986).
4. Morelli, A. & Dziewonski, A. M. *Nature* **325**, 678–683 (1987).
5. Poupinet, G., Pillet, R. & Souriau, A. *Nature* **305**, 204–206 (1983).
6. Morelli, A., Dziewonski, A. M. & Woodhouse, J. H. *Geophys. Res. Lett.* **13**, 1545–1548 (1986).

7. Ritzwoller, M., Masters, G. & Gilbert, F. *J. geophys. Res.* **91**, 10203–10228 (1986).
8. Woodhouse, J. H., Giardini, D. & Li, X.-D. *Geophys. Res. Lett.* **13**, 1549–1552 (1986).
9. Giardini, D., Li, X.-D. & Woodhouse, J. H. *Nature* **325**, 405–411 (1987).
10. Dziewonski, A. M. & Anderson, D. L. *Phys. Earth planet. Inter.* **25**, 297–356 (1981).
11. Toy, K. M., Olson, A. H. & Orcutt, J. A. *Trans. Am. geophys. Un.* **67**, 1099 (1986).
12. Gudmundsson, O., Clayton, R. W. & Anderson, D. L. *Trans. Am. geophys. Un.* **68**, 1378 (1987).
13. Backus, G. E. *J. geophys. Res.* **70**, 3429–3439 (1965).
14. Crampin, S. *Geophys. J. R. astr. Soc.* **49**, 9–27 (1977).

Crystal structure of an antifreeze polypeptide and its mechanistic implications

D. S. C. Yang*, M. Sax*, A. Chakrabartty† & C. L. Hew†

* Biocrystallography Laboratory, VA Medical Center, PO Box 12055, Pittsburgh, Pennsylvania 15240 and Department of Crystallography, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

† Research Institute, Hospital for Sick Children, Toronto and Departments of Biochemistry and Clinical Biochemistry, University of Toronto, Toronto M5G 1L5, Canada

The X-ray crystallographic structure of an antifreeze polypeptide from the fish winter flounder, has been determined at 2.5 Å by an analysis of the Patterson function. This is the first report of a polypeptide of this size that is a single α -helix. A proposed mechanism of antifreeze binding to ice surfaces is given which requires: first, that the dipole moment from the helical structure dictates the preferential alignment of the peptide to the c-axis of ice nuclei; second, amphiphilicity of the helix; and third, torsional freedom of the side chains to facilitate hydrogen bonding to ice surfaces.

FREEZING is lethal to most cellular organisms. Dehydration of the intracellular environment and physical damage by ice crystals are the major causes for freezing injury and death. Many marine fishes living in cold, polar waters survive in the ice-laden surroundings by synthesizing a unique class of serum proteins which prevent their blood sera from freezing. Aqueous solutions containing these proteins, commonly known as antifreeze proteins, possess many unusual freezing properties^{1–3}.

Isolation and subsequent characterization of these proteins from different fishes have established the diversity of their chemical structures. At least one type of antifreeze glycoprotein (AFGP) and three types of antifreeze polypeptides (AFP) are known. All the AFGPs isolated so far are similar in chemical structure. They are polymers of a glycotriptide unit, Ala Ala Thr, with a disaccharide linked to the Thr side chain. The AFPs, on the other hand, are quite diverse structurally. They can be grouped into three classes, namely, the alanine-rich AFPs, cysteine-rich AFPs and a third class that is neither rich in alanine nor cysteine⁴.

The properties of antifreeze proteins have been investigated using a number of techniques including freezing point osmometry, ¹³C-NMR, electron microscopy, Raman spectroscopy, light scattering and so on^{3,5}. All AFGPs and AFPs appear

to operate via a noncolligative mechanism and depress freezing temperature to a greater extent than the melting temperature. On a weight basis they are as efficient as salt (NaCl) in lowering the freezing temperature, but on a molar basis they are 300–500-fold more efficient.

We report here the first crystallographic analysis of the three dimensional structure of an antifreeze polypeptide molecule. The initial phase determination was based on an analysis of the Patterson function, therefore, it is also the first report of a structure of this size that was essentially determined by the Patterson method. The antifreeze molecule is a single α -helix. Based on the well established dipolar character of α -helix^{6,7}, a mechanism has been put forth to explain the specific binding of antifreeze molecules to the prism faces (defined as faces parallel to the c-axis) of the ice nuclei.

Alanine-rich AFPs

This class of AFP is the most extensively characterized. Those that have been studied were isolated from fish from two different phylogenetic families; the winter flounder of the right eye flounder family, and the shorthorn, arctic and grubby sculpins of the cottid family^{8–10}. The amino-acid sequences of these AFPs are shown in Table 1. The structural characteristics of these AFPs

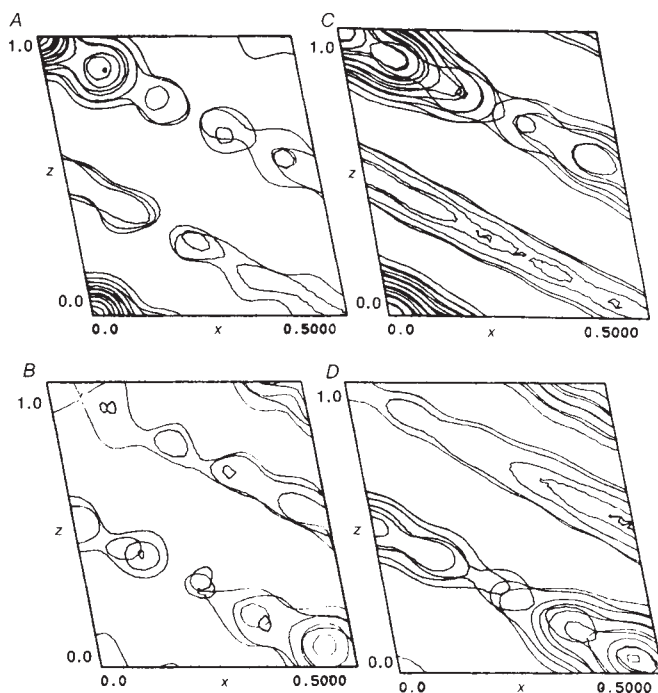


Fig. 1 Patterson maps. A, native Patterson at $V=0.0$; B, native Patterson at $V=0.5$; C, model Patterson at $V=0.0$; D, model Patterson at $V=0.5$.

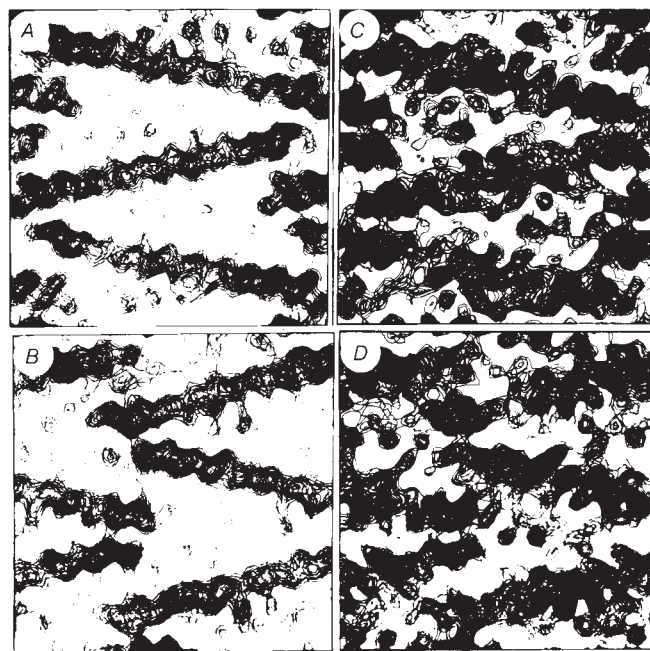


Fig. 2 Electron density maps at 4.5 Å resolution. The screw axes are vertical. A, B, molecules one and two phased by both helices; C, D, molecules one and two phased by one helix.

are their relatively small relative molecular masses (M_r , 3,300–4,500), the multiplicity of active components, the high alanine content (up to 60 mol %), and the high α -helical structure (especially at low temperatures, as estimated by circular dichroism studies¹¹). In addition, many AFPs contain an 11-amino-acid repeat consisting of Thr- X_2 -polar amino acid- X_7 , where X is mainly alanine. For example, each of the two major AFPs from the winter flounder contains three of these repeats. Another common feature of these AFPs is that if they are folded into an α -helix most of their hydrophilic side chains would fall on one side of the helix.

4.5 Å structure

The major winter flounder AFP (HPLC-6) was selected for the crystal structure analysis. Crystals grown from acetone were mounted in glass capillaries for use in measuring the diffracted

intensities on a Nonius CAD-4 diffractometer operating at 45 kV and 26 mA with Cu K α radiation. The crystals belong to space group $P2_1$ with cell parameters $a = 38.0$, $b = 37.06$, $c = 21.75$ Å and $\beta = 101.45^\circ$ (ref. 12). There are two molecules of 3,300 daltons per asymmetric unit. Data collection at room temperature was hampered by condensation around the crystal which caused crystal slippage. This problem was alleviated by soaking the crystals in isopropanol. With the new solvent system, the crystals stayed stationary for two days before slippage occurred. During the stable period, a set of 4.5 Å data were collected employing ω -scans with a rate of $0.3^\circ \text{ min}^{-1}$. Background for each reflection was determined by first fitting a Lorentzian/Gaussian curve to the intensity profile (96 steps) to determine the peak position and the background cutoff point, and then fitting a least squares line to the background region. The net intensity was determined by a stepwise subtraction of the

Table 1 Amino-acid sequences of alanine-rich AFP

I. Righteye flounders

Winter flounder	1	10	20	30	$\langle\mu\rangle$	$\langle H\rangle$
HPLC-6	DTASDAAAAAALTAANAKAAAEELTAANAAAAAATAR				0.13	0.21
HPLC-8	DTASDAAAAAALTAANAAAAAKLTADNAAAAAATAR				0.19	0.21

II. Cottids

Shorthorn sculpin	1	5	10	20	30	$\langle\mu\rangle$	$\langle H\rangle$
SS-3	MNAPARAAAKTAADALAAAKKTAADAAAAA					0.27	0.16
SS-8	1	10	20	30	40		
Grubby sculpin	MNGETPAQKAARLAAAAALAAKTAADAAAKAAKAAIAAAAA					0.26	0.20
GS-5	1	5	10	20	30		
Arctic sculpin	MDAPAIAAAKTAADALAAAKKTAADAAAAA					0.24	0.19
AS-1	1	5	10	20	30		
	MDAPARAAAKTAADALAAANKTAADAAAAA					0.27	0.17
AS-3	1	10	20	30			
	MDGETPAQKAARKAAAAAALAKTAADAAAAA					0.24	0.09

$\langle\mu\rangle$, mean hydrophobic moment calculated according to ref. 23 using the 'normalized consensus' (hydrophobicity scale from ref. 23). $\langle H\rangle$, mean hydrophobicity, also calculated according to ref. 23. Amino-acid sequences are from refs 2, 3, 8–10.

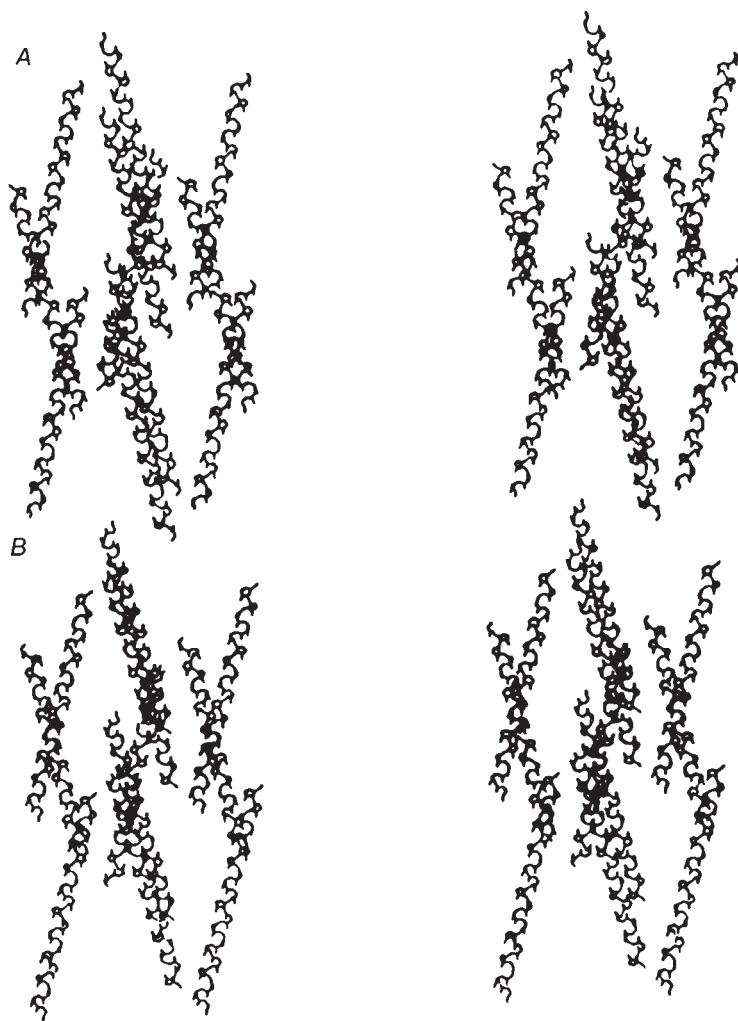


Fig. 3 Stereo packing diagrams of the two helices. The screw axes are horizontal. A, Helix one is the centre of the diagram; B, helix two in the centre of the diagram.

background from the peak value. An L_p and an empirical absorption correction were made. Radiation decay was monitored by measuring six standard reflections at 4-h intervals during the data collection. No observable decrease in the intensity of the standard reflections occurred. The merging R_{mer} on $|F|^2$ for the 29 duplicates among 420 reflections was 3%. Among the 391 unique data measured, 81% had intensities exceeding $4\sigma F$.

The determination of the crystal structure is interesting in that it was based mainly upon a direct analysis of the Patterson map¹³ of the native protein, which is unusual for a polypeptide of this size. Circular dichroism studies¹¹ indicated that the AFP possesses a high α -helical content. For this element of secondary structure, it is known that the parallel interatomic vectors appear in the Patterson function as a chain of peaks, parallel to the helix axis with an interpeak spacing equal to the distance of the helical periodicity. These parameters for a single helix of periodicity 5.4 Å were easily identified in the Patterson map (Fig. 1A, B). Attempts to find similar features for the other helix in the asymmetric unit were not successful, unless the two helices are considered to be parallel. This arrangement of the helices is consistent with the salient features of the Patterson map. It explains a chain of peaks centred at $U=0.5$, $V=0.5$, $W=0$ and separated by 5.4 Å, which is parallel to the chain at the origin, as representing the vectors between the two parallel helices. The remaining peaks at $V=0.0$ and $V=0.5$ are those relating one helix to the symmetry-related molecule of the other helix. These vectors limit the allowed positions of the helices. It is clear that each helix intercepts a different screw axis.

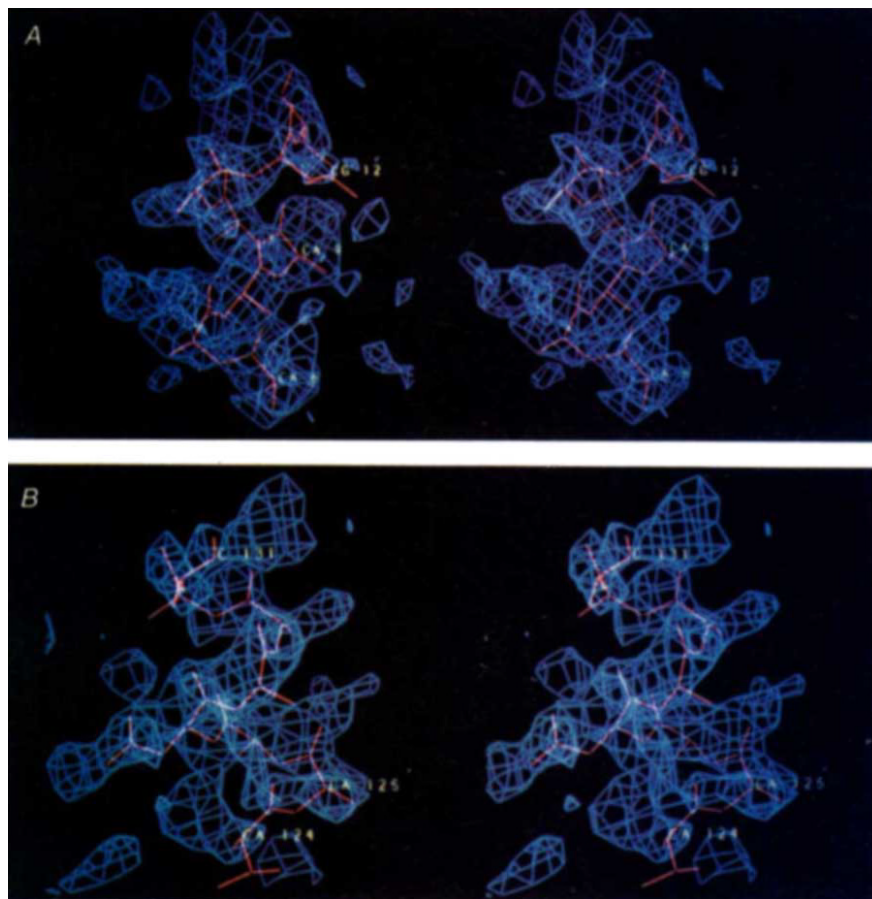
There remained only three parameters to be determined for a solution of the structure at this resolution. These were: first,

translation perpendicular to the screw axis in the plane containing the helix and the screw axis; second, rotation about the two helical axes; and third, the sense of direction of the two helical axes.

Two model α -helices with 36 residues of glycine were used in an R -factor translation and rotation search which was performed on a 4.7 Å grid with 60° rotation intervals. The search was coupled with ten cycles of rigid body refinement at each point which allows six degrees of freedom for each helix (programme GRPREF; W. Furey, unpublished results). The temperature factors were fixed at 15.0 Å². The geometry of the helices was monitored with the programme LSLINE¹⁴ which calculates the interhelical angle and distance as an additional guide for assessing the stability of the refinement. One unique solution was found with $R=47.6\%$ and the next lowest was 49%. A finer interval of 1.9 Å and 30° was then used for another search which yielded a solution with $R=47.1\%$. The resulting interhelical angle was 28° and distance was 8.1 Å. The inclinations of the two helices to the x - z plane were 12.7° and 15°. The Patterson function corresponding to this final refined model looked almost identical to that generated from experimental data (Fig. 1C, D). Further refinement with a fixed temperature factor of 5.5 Å² gave an R factor of 43.0%.

The refined model of the helices was used to generate phases for an electron density map calculation. Interestingly, some details that are not usually observable at 4.5 Å resolution are evident in this map (Fig. 2A, B), such as the salt bridge between Lys₁₈ and Glu₂₂. As the model contains only polyglycine, it gave no information on the side chains. To further demonstrate that the map is not an artifact produced by the model, one helix was omitted from the model for phase calculation. A Fourier map

Fig. 4 Residue deleted $2|F_0| - |F_c|$ electron density maps at 2.5 Å resolution contoured at 0.275 of the maximum density. Coordinates are from the final refinement cycle ($R = 27.4\%$, 1,085 reflections $> 10.0 \sigma F$, 5.0–2.5 Å, overall temperature factor). **A**, Residues 7–12 deleted from phase calculation; **B**, Residues 125–130 deleted from phase calculation.



computed with these phases contained the omitted helix in the appropriate location and with electron density values at approximately half those of the phasing helix (Fig. 2C, D).

An analysis of the packing arrangement of the helices indicated that each is in contact with six other molecules (Fig. 3A, B). The modes of their interaction with neighbouring helices, though not identical, are quite similar. Each makes head-to-tail contacts with symmetry-related molecules. In addition, it is criss-crossed by four nonsymmetry-related molecules, two on each side. The helices thus form a very stable network.

2.5 Å structure

A complete set of 2.5 Å data was collected from one crystal on a CAD4 diffractometer at 0 °C. Lowering of the temperature solved the crystal slippage problem. There is no decrease in the intensity of the six standard reflections during the data collection. Among the 2,109 unique data measured, 88% had intensities exceeding $4\sigma F$. The data scaled well with the previous set of data with $R_{\text{mer}} = 6.6\%$. The model helices this time consisted of 36 residues of alanine and were subjected to restrained least-squares refinement^{15,16} using the new data set. Temperature factors were refined isotropically. The refinement converged to a conventional R of 36%. At this point the coordinates were examined using an Evans and Sutherland graphic system. All side chains were generated and positioned in a $2|F_0| - |F_c|$ map. Upon subsequent refinement the R -value converged to 31.8% (1,512 reflections $> 6.0\sigma F$, 8.0–2.5 Å, individual temperature factors). At this stage residue-deleted maps were generated by omitting six residues from the model at a time in the phase calculation. The map regions corresponding to helix one were much cleaner than regions corresponding to helix two. The moderately high R -value (31.8%), the noisy maps and the fact that the two refined helices were parallel rather than anti-parallel prompted us to suspect that the direction of helix two might have been in error. Due to the high content of alanine and other short side chains the searches performed at 4.5 Å could possibly

have picked up a solution with one of the helices having its direction reversed. We therefore refitted the density map with helix two inverted (with a polyalanine backbone) and performed a series of rigid-body refinements. The R -value converged to 32.3% and subsequent refinement with all side chains added converged to 27.4% (1,085 reflections $> 10.0 \sigma F$, 5.0–2.5 Å, overall temperature factor). Some of the residues are shown illustrating the fit (Fig. 4A, B). The electron densities of some of the side chains are not very well defined indicating some sort of disorder, especially at the terminal sites. This observation is not surprising as it has been demonstrated that residues with discrete multiple conformations are frequently found in protein crystals on molecular surfaces¹⁷. Considering the fact that all side chains of the antifreeze are situated on the molecular surface it is reasonable to assume that they might have discrete multiple conformations. Due to the moderate resolution of the data no attempts have been made to analyse the disorder in the side chains nor to locate the solvent molecules.

Mechanism of action

It is well documented that, in the absence of antifreeze protein, ice nuclei grow perpendicular to their c -axes thus enlarging the basal plane¹⁸. Addition of antifreeze protein retards or stops the growth of the basal plane and, as the protein concentration increases, the preferred growth direction shifts to become parallel to the c -axis³. Raymond and DeVries¹⁹ postulated that antifreeze protein might bind preferentially to faces of the ice nucleus that are parallel to the c -axis (prism faces) and subsequently inhibit growth of the basal plane. They attribute the inhibitory activity to adsorption of the protein to the ice nuclei, and they assert that presence of the protein depresses the freezing temperature by raising the curvatures of surfaces on the ice crystal.

The Raymond-DeVries model accounts for the inhibition of growth and depression of the freezing temperature of ice crystals in a plausible manner, but it does not explain why AFP molecules

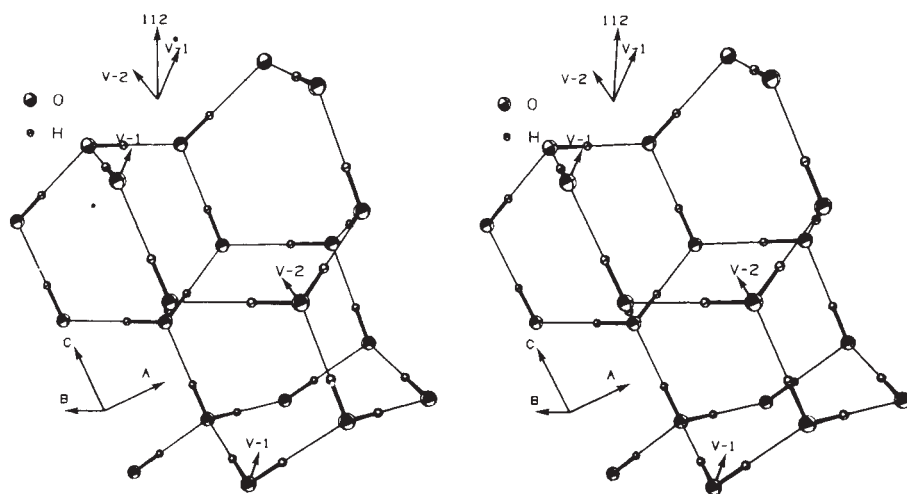


Fig. 5 Stereo diagram of an ice lattice illustrating various vectors mentioned in the text.

adsorb preferentially to prism faces. Assuming the AFP molecules are α -helices, DeVries and Lin²⁰ and DeVries¹ attribute this directional preference to matching of the distance between threonine and aspartate groups of an AFP to the distance between water molecules on the prism faces. Why the protein and the ice crystal should align in this specific orientation remains obscure, however, because compatible matches also occur between water molecules on the basal plane and the polar side-chain groups of the protein. This is clear from the crystal structure of the AFP which shows that distances between polar side-chain groups can and do assume different values, including some of those contained in different faces of ice (for example, 2.8, 4.5, 4.6 Å, and so on).

Our crystallographic results show that the flounder antifreeze molecule is a single α -helix and that due to the alignment of the peptide dipoles⁶, the AFP possesses a significant dipole moment parallel to its helix axis. We propose that interaction of the AFP dipole with those of the water molecules in the bound ice nucleus may provide a better explanation for the preferential orientation of the helix with respect to the ice lattice. Dipole-dipole interactions of this kind may be a requirement for inhibiting the growth of ice nuclei. According to this hypothesis, the orientations of dipoles on the water molecules in the ice nuclei are expected to align themselves preferentially in a direction antiparallel to the adjacent helix dipole. This would reduce disorder in the hydrogen positions of the ice. In a perfectly ordered ice structure, the orientation of the dipole within a particular layer normal to the c -axis will be inclined by 54.7° to the c -axis and 30° to the a -axis (reciprocal vector $V-1$, Fig. 5). The orientation of the dipole in the adjacent layer will be inclined by 54.7° to the c -axis and 90° to the a -axis (reciprocal vector $V-2$) however. The resultant (reciprocal vector $\langle 112 \rangle$) of these two vectors is inclined 50.8° to the c -axis and 60° to the a -axis. Consequently, maximum interaction between the helix dipole and the local ice dipoles occurs when the helical axis is parallel to the $\langle 112 \rangle$ vectors.

In sufficiently dilute solutions, the concentration of AFP on the surface of ice nuclei is too low for significant helix-helix interactions. Helix-water interactions dominate, and the helices lie on the prism faces of the ice nuclei with their axes parallel to the $\langle 112 \rangle$ vectors. This interaction between AFP and the ice surface induces local ordering of the water molecules on the outer layers of the prism faces and retards growth of the basal plane. Further growth of the ice nuclei, in the c -axis direction, is relatively unaffected at locations away from the ordered outer layers. This is because addition of water molecules to an ordered ice lattice is entropically less favourable than to an disordered ice lattice. This growth habit should result in bipyramidal ice crystals with steps parallel to the c -axis, as indeed observed in the microscope (Fig. 6A,D). This mechanism is consistent with the result reported by Knight *et al.*²¹ who observed the develop-

ment of $\{10\bar{1}x\}$ faces of ice in the presence of low concentration of AFP. With higher concentrations of AFP, the helix-helix dipole interactions on each prism face should increase, causing the helices to become antiparallel. Subsequent interactions between helices on different prism faces cause their helical axes to gradually become parallel to the c -axis of the ice crystal. This altered growth habit should increase the step size of the bipyramidal ice crystals, and was actually detected (Fig. 6B,E). With sufficiently high concentrations of AFP, all helices are parallel to the c -axis. Thus, ordering of the hydrogen positions of the ice crystal should only be along the x -axis with little penetration beyond the surface layers. Therefore, the step size

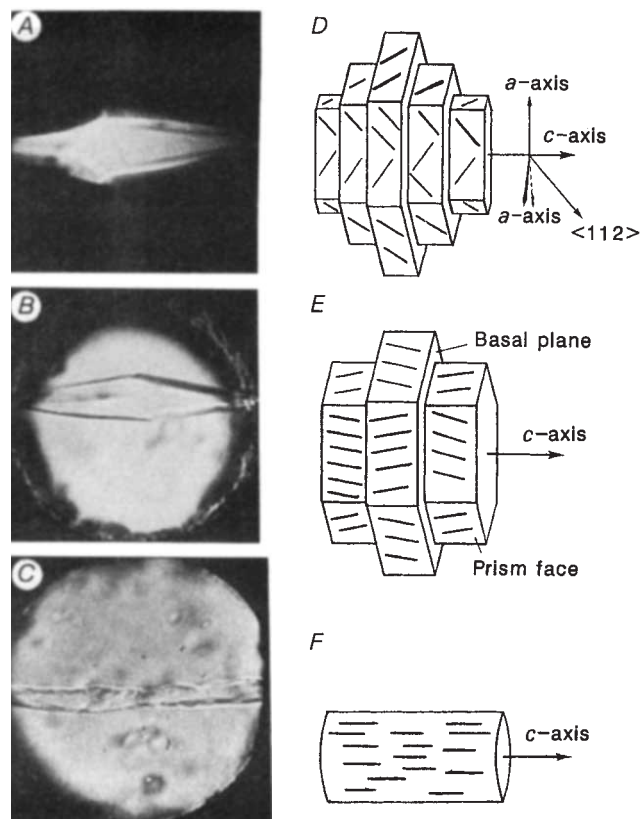


Fig. 6 Photomicrographs and schematic diagrams of ice crystals formed in solutions of different protein concentrations using an advanced nanolitre osmometer. A, AFP concentration of 0.242 mM, using cross-polarized light; B, AFP concentration of 0.363 mM; C, AFP concentration of 3.63 mM; D, E and F, schematic diagrams of A, B and C, respectively. The c -axes of each crystal are horizontal.

approaches infinity and the bipyramidal shape becomes needle-like (Fig. 6C,F).

The dipole interactions between the AFP and water molecules in the ice lattice provide a reasonable explanation for the preferential alignment with ice nuclei. It can be envisioned that the next requirement for binding would be amphiphilicity of the helices. It should be noted that the AFPs are not strictly amphiphilic in the traditional sense because there are non-hydrophilic residues on the 'hydrophilic side' of the helices, most of these being alanines. As a consequence, the hydrophobic moments of the AFP helices are rather small ($\langle\mu\rangle$ ranges from 0.13 to 0.27, Table 1 and ref. 22). They are at least partially amphiphilic, however, because most of their hydrophilic side chains do fall on one side of the helices. It is reasonable to assume that while the hydrophilic side chains in the polypeptides are required for hydrogen bonding with the ice lattice, the cluster of hydrophobic side chains will impede further growth of the ice nuclei. The 11-amino-acid repeat observed in many of these alanine-rich AFPs corresponds to three helical turns and is presumably for generating the amphiphilicity in the helices.

According to the above mechanism, the AFPs bind to prism faces of ice in different orientations at different protein concentrations. Therefore, it is important that the protein's hydrophilic side chains have enough torsional freedom to accommodate all of these interactions. These torsional variations are observed in the crystal structure of the flounder AFP. The hydrophilic side chains of the other alanine rich AFPs are also expected to have such degrees of freedom.

Conclusions

In the beginning of our attempt to formulate a mechanism of

antifreeze binding, we concentrated on matching the distances between protein hydrophilic side chains and the distances separating oxygen atoms on the various planes of the ice crystal lattice. It was finally realized that side chains on protein surfaces do adopt discrete multiple conformations; therefore, such a static match does not exist between the protein and ice surfaces. On the contrary, as all surfaces of the ice crystal are densely populated with atoms that can hydrogen bond to the protein surface, the flexibility of the side chains means that many patterns of hydrogen bonding can exist. It was realized later that the dipolar character of α -helix might be an important factor in this binding process. Initially, by random diffusion an AFP molecule encounters an ice nucleus and induces local ordering of the ice lattice. Then the electrostatic interactions between the helix and induced dipoles align the AFP with respect to the ice lattice. After the alignment, subtle shifts of the helical conformation and torsional movement in the side chains of the hydrophilic amino acids help to complex the AFP with the ice surface.

The mechanism of antifreeze binding presented in this paper may be a general one and applicable to other alanine-rich AFPs. It is not yet clear, however, how the mechanism can be applied to the other antifreeze proteins. It is very likely that the answer can only be derived from the three-dimensional structures of these proteins.

We thank W. Furey and R. Love for instructions on using the refinement programs and graphic system; J. Rosenberg and D. Abraham for use of their picture systems; D. B. Huang for assistance and E. Whalley, J. Nagle and S. B. Joshi for discussions. The work was supported by a CATI Grant from the University of Pittsburgh and by grants from NIH and MRC Canada.

Received 12 November 1987; accepted 7 April 1988.

- DeVries, A. L. *Phil. Trans. R. Soc. B* **304**, 575-588 (1984).
- Hew, C. L. & Fletcher, G. L. in *Circulation, Respiration and Metabolism* (ed. Gilles, R.) 553-563 (Springer, Berlin, 1985).
- Feeney, R. E., Burcham, T. S. & Yeh, Y. *Ann. Rev. Biophys. biophys. Chem.* **15**, 59-78 (1986).
- Hew, C. L., Scott, G. K. & Davis, P. L. in *Living in the Cold, Physiological and Biochemical Adaptation* (eds Heller, H. D. et al.) 117-123 (Elsevier, New York, 1986).
- Berman, E., Allerhand, A. & DeVries, A. L. *J. biol. Chem.* **255**, 4407-4410 (1980).
- Hol, W. G. J., van Duijnen, P. T. & Berendsen, H. J. C. *Nature* **273**, 443-446 (1978).
- Shoemaker, K. R. et al. *Nature* **326**, 563-567 (1987).
- Davies, P. L., Roach, A. H. & Hew, C. L. *Proc. natn. Acad. Sci. U.S.A.* **79**, 335-339 (1982).
- Pickett, M. H. et al. *Eur. J. Biochem.* **143**, 35-38 (1985).
- Chakrabarty, A. et al. *Can. J. Zool.* (in the press).

- Ananthanarayanan, V. S. & Hew, C. L. *Biochem. biophys. Res. Commun.* **74**, 685-689 (1977).
- Yang, D. S. C., Chung, Y. J., Chen, P., Rose, J. P. & Hew, C. L. *J. molec. Biol.* **189**, 725-725 (1986).
- Patterson, A. L. *Z. Kristallogr.* **90**, 517-521 (1935).
- Yang, D. S. C. thesis, Univ. Pittsburgh (1983).
- Furey, W. et al. *J. appl. Crystallogr.* **15**, 160-166 (1982).
- Hendrickson, W. A. & Konnert, J. H. *Biomolec. Struct. Funct. Conform. Evol.* Vol. I, 43-57 (1981).
- Svensson, L. A. et al. *Proteins: Struct. Funct. Genet.* **1**, 370-375 (1986).
- Fletcher, N. H. *The Chemical Physics of Ice* (Cambridge University Press, 1970).
- Raymond, J. A. & DeVries, A. L. *Proc. natn. Acad. Sci. U.S.A.* **74**, 2589-2593 (1977).
- DeVries, A. L. & Lin, Y. *Biochim. biophys. Acta* **495**, 388-392 (1977).
- Knight, C. A., DeVries, A. L. & Oolman, L. D. *Nature* **308**, 295-296 (1984).
- Eisenberg, D., Weiss, D. & Terwilliger, T. C. *Nature* **299**, 371-374 (1982).
- Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, E. *J. molec. Biol.* **179**, 125-142 (1984).

LETTERS TO NATURE

A millisecond pulsar in an eclipsing binary

A. S. Fruchter, D. R. Stinebring & J. H. Taylor

Joseph Henry Laboratories and Physics Department,
Princeton University, Princeton, New Jersey 08544, USA

We have discovered a remarkable pulsar with period 1.6 ms, moving in a nearly circular 9.17-h orbit around a low-mass companion star. At an observing frequency of 430 MHz, the pulsar, PSR1957+20, is eclipsed once each orbit for about 50 minutes. For a few minutes before an eclipse becomes complete, and for more than 20 minutes after the signal reappears, the pulses are delayed by as much as several hundred microseconds—presumably as a result of propagation through plasma surrounding the companion. The pulsar's orbit about the system barycentre has a radius of 0.089 light seconds projected on to the line of sight. The observed orbital period and size, together with the fact that eclipses occur, imply a surprisingly low companion mass, only a few per cent of the mass of the Sun.

The new pulsar was discovered in the course of a general

survey for millisecond pulsars being carried out with the 305-m telescope of the Arecibo Observatory. At 430 MHz this antenna provides sensitivity of $8\text{--}18\text{ K Jy}^{-1}$, depending on zenith angle, and a typical system temperature in the direction of the galactic equator of $\sim 150\text{ K}$. Our survey observations were directed at a grid of positions separated by 10 arc min, covering about 250 square degrees centred on the galactic plane in the region from $40^\circ\text{--}60^\circ$ galactic longitude. Each position was observed for 67.7 s, and the output of a 128-channel autocorrelation spectrometer observing the band 425–435 MHz was recorded every 516.625 μs .

The survey observations were analysed on a Cyber 205 at the John von Neumann Center in Princeton. The autocorrelation data were Fourier transformed to produce 128-point spectra for each of the 131,072 sample intervals recorded for a single beam area. The resulting two-dimensional (time and frequency) arrays were then combined at 512 trial dispersion measures over the range $0\text{--}254\text{ cm}^{-3}\text{ pc}$, Fourier transformed in the long (time) dimension, and searched thoroughly for the presence of periodic signals. The survey sensitivity is 1 mJy for pulsars with period $P > 15\text{ ms}$ and dispersion measure $\text{DM} < 60\text{ cm}^{-3}\text{ pc}$, increasing to about 10 mJy at $P = 1.5\text{ ms}$ and $\text{DM} = 120\text{ cm}^{-3}\text{ pc}$. Complete details of the searching procedure and further results of the survey will be published elsewhere.

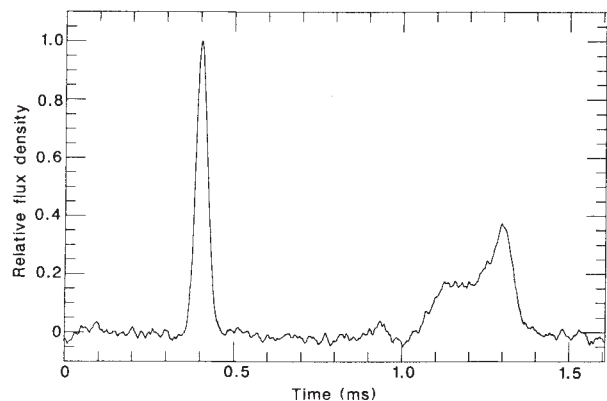


Fig. 1 Integrated profile of PSR1957+20 at 430 MHz. Instrumental smoothing amounts to $\approx 15 \mu\text{s}$, mostly a result of drifting pulse phase during the integration caused by imperfect knowledge of the period.

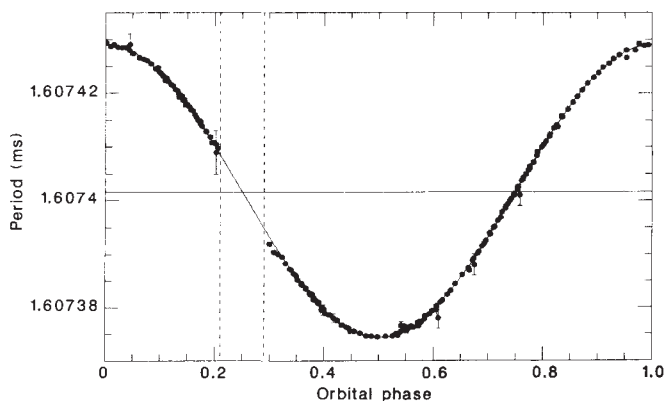


Fig. 2 Orbital velocity curve of PSR1957+20. The pulsar is eclipsed by its companion between phases 0.21–0.29. Points with visible error bars were taken using the pulsar survey system; all other points were obtained with the data acquisition system designed for timing millisecond pulsars.

The signal from PSR1957+20 was first observed in October 1986, during some of the earliest observations of the search, but re-observation of the accumulated list of candidate pulsars was not begun until March 1988. The existence of PSR1957+20 was confirmed on 3 March 1988; observations over the next several days showed its period to be variable at a level corresponding to orbital velocities of $\pm 5 \text{ km s}^{-1}$. These observations also revealed the presence of a strong interpulse. On 5 March, at about 13:30 UT, the signal disappeared abruptly for no apparent reason. As described below, we now realize that the pulsar went into eclipse behind its companion star.

Additional observations were made for about 2.5 hours per day on 9 days between 9 and 28 March. These measurements made use of two data acquisition systems explicitly designed for timing millisecond pulsars^{1,2}. The newest system, used for most of the observations, coherently de-disperses the signals received in two 0.4-MHz passbands centred at 427 and 430 MHz. After square-law detection, these signals are sampled, integrated synchronously with the topocentric pulsar period, and recorded about once a minute. The sum of 20 of these integrated profiles is plotted in Fig. 1, illustrating the main pulse of width $30 \mu\text{s}$ and a much wider interpulse separated by about half a period.

An equivalent pulse arrival time was determined for each integration by a process of template fitting. Analysed in short

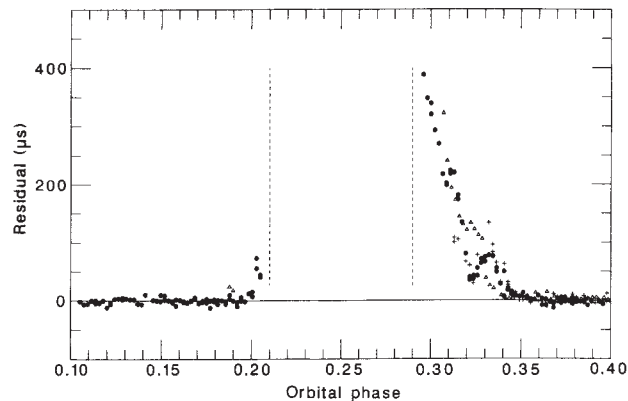


Fig. 3 Excess group delays in the pulsar signal at 430 MHz relative to a model based on our best-fit parameters, plotted as a function of orbital phase near the centre of eclipse. Different symbols correspond to different observing days.

segments of $\sim 5 \text{ min}$ duration, the arrival times (reduced to the solar system barycentre) yield periods which, when folded modulo the 9.17 hour orbital period, comprise the velocity curve illustrated in Fig. 2. A linearized least-squares fit to the period data yielded the pulsar and orbital parameters quoted in Table 1. A 'phase-connected' solution to the observed pulse arrival times, with no pulse-numbering ambiguities between observing days, should be achievable soon and will provide much more accurate pulsar and orbital parameters.

The entire orbit has now been observed. Good observations exist on five different days for all or part of the region $0.15 < \phi < 0.35$, where orbital phase ϕ is measured from the time of the ascending node. On all of these days the pulsar signal has disappeared completely between the orbital phases 0.21 and 0.29. At $\phi = 0.25$, the pulsar is at its greatest distance from Earth, and closest to being directly behind the companion star. We conclude that the intervals of missing signal are the result of eclipses by a surprisingly large, though very low-mass, companion.

During the few minutes immediately before eclipse, and for $\sim 20 \text{ min}$ after the eclipse, the pulsar signal is delayed by as much as $400 \mu\text{s}$ relative to arrival times predicted from a model fitted to parts of the orbit farther from eclipse. These effects are illustrated in Fig. 3, which shows that although the eclipse interval is very nearly symmetric about $\phi = 0.25$, the magnitude and duration of the excess signal delays are far from symmetric, and appear to change significantly from orbit to orbit.

The excess delays are probably the result of propagation in ionized gas surrounding the companion. The maximum delay, just after the pulsar has emerged from eclipse, corresponds to an increase in dispersion measure of $\sim 0.017 \text{ cm}^{-3} \text{ pc}$, or equivalently an electron column density of $5 \times 10^{16} \text{ cm}^{-2}$. Unfortunately our two frequency channels at 427 and 430 MHz are too closely spaced to determine reliably the frequency dependence of the effect. Away from eclipse the dispersion measure to the pulsar is found to be $29.13 \pm 0.01 \text{ cm}^{-3} \text{ pc}$.

Table 1 Parameters of the PSR1957+20 system

Right ascension (1950.0)	$\alpha = 19 \text{ h } 57 \text{ min } 10 \text{ s} \pm 20 \text{ s}$
Declination (1950.0)	$\delta = 20^\circ 40' \pm 5'$
Pulsar period	$P = 1,607.40171 \pm 0.00003 \mu\text{s}$
Dispersion measure	$DM = 29.13 \pm 0.01 \text{ cm}^{-3} \text{ pc}$
Flux density at 430 MHz	$S = 25 \pm 10 \text{ mJy}$
Projected semi-major axis	$a_1 \sin i = 0.08923 \pm 0.00007 \text{ light s}$
Eccentricity	$e < 0.001$
Orbital period	$P_b = 33,001.9 \pm 0.5 \text{ s}$
Time of ascending node	$T_0 = 2,447,245.08471 \pm 0.00004 \text{ JED}$

The orbital parameters listed in Table 1 correspond to a pulsar mass function

$$\left(\frac{2\pi}{P_b}\right)^2 \frac{(\alpha_1 \sin i)^3}{G} = \frac{(m_2 \sin i)^3}{(m_1 + m_2)^2} = (5.20 \pm 0.01) \times 10^{-6} M_\odot$$

where G is the gravitational constant, m_1 and m_2 are the masses of the pulsar and companion, and i is the inclination between the plane of the orbit and the plane of the sky. The observed eclipses suggest that $\sin i$ is not much less than 1.0, so for a pulsar mass $m_1 \approx 1.4 M_\odot$ the companion mass must be near the minimum possible mass of $0.022 M_\odot$. Indeed, even without the observation of eclipses, the *a priori* probability of finding $\sin i < 0.4$, and hence $m_2 > 0.055$, is less than 10%. A pulsar mass of $1.4 M_\odot$ further implies that the ratio m_1/m_2 is around 60, and that the radius of the companion's orbit is near $2.4 R_\odot$. As the eclipse lasts for about 10% of an orbit, we find that the opaque portion of the companion star must be at least $1.5 R_\odot$ across.

The radius of the Roche lobe of the companion is only $\sim 0.3 R_\odot$, so most of the volume occupied by the eclipsing body lies well outside it. We suggest that the eclipsing plasma may be a stellar wind, powered by the ~ 100 solar luminosities of energy that would be emitted by the pulsar were it to spin down at a rate comparable to those of other millisecond pulsars. The slow decrease of plasma density on the trailing side of eclipse might then be interpreted as a comet-like tail spewing off the companion.

Indeed, it is exciting to speculate that we are witnessing the evaporation of the companion by the pulsar. Although it is difficult to reconstruct the density of the plasma without knowing the geometry of the object, the rapid variation in the observed column density suggests a scale size perhaps as small as 0.1 light seconds, and thus a plasma density exceeding 10^7 cm^{-3} in the transparent region. If the plasma concentration continues to rise steeply in the region that is now obscured, and if the wind reaches velocities comparable to the companion's orbital velocity, it seems possible that the companion will disappear in much less than 10^6 years, leaving behind an isolated millisecond pulsar. We expect that, in the near future, radio observations at higher frequencies, and perhaps spectroscopy of the wind, will allow us to determine whether there is any truth in this fascinating possibility.

We thank G. Berman, M. Davis, J. Hagen, T. Hankins, P. Perillat and M. Ryba for essential contributions to observational work, and B. Paczynski and C. Thompson for valuable discussions concerning binary evolution. This work was made possible by the facilities of the National Astronomy and Ionosphere Center and the John von Neumann National Supercomputer Center, and was supported by grants from the US National Science Foundation and the National Bureau of Standards. A.S.F. is supported by a John von Neumann Fellowship.

Received 1 April; accepted 22 April 1988.

1. Rawley, L. A., Taylor, J. H., Davis, M. M. & Allen, D. W. *Science* **238**, 761–765 (1987).
2. Hankins, T. H., Stinebring, D. R. & Rawley, L. A. *Astrophys. J.* **315**, 149–153 (1987).

A supernova in the nucleus of NGC5548?

Roberto Terlevich* & Jorge Melnick†

* Royal Greenwich Observatory, Herstmonceux Castle, Hailsham, East Sussex BN27 1RP, UK

† European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-8046 Garching, München, FRG

Peterson and Ferland¹ have reported observations of the low-luminosity variable Seyfert galaxy NGC5548, which they interpret as the first direct evidence for an accretion event in the nucleus of an active galaxy. Here we show that the spectrum, the timescale and the luminosity of the event may be better interpreted as the first direct evidence for a supernova explosion in the nuclear region of a spiral galaxy.

The energy source most frequently adduced for active galactic nuclei (AGN) is accretion onto a massive collapsed object, often called 'the monster'^{2–4}. We have challenged this view on the grounds that most of the observations usually invoked as evidence for the existence of nuclear monsters can also be understood in terms of violent star formation activity^{5,6}: the observational diagnostics (emission line ratios, radio and X-ray emission, and so on) which have traditionally been used to distinguish between 'normal' (starburst) and 'active' emission-line nuclei do not distinguish between starbursts and monsters⁶.

A promising observational approach for probing the innermost regions of AGNs is the study of their variability. In 'monster' models, variability is ascribed to changes in the accretion rates, whereas in the starburst theory, variability is the consequence of transient phenomena associated with the evolution of massive stars and in particular with their explosions as supernovae and subsequent evolution as (young) supernova remnants in a high-density interstellar medium^{6,7}. Long-term monitoring of AGNs has been carried out over the past few years for a few galaxies, and has already yielded interesting

results^{7–13}. We have shown elsewhere that these results support the starburst scenario⁶.

During the monitoring of NGC5548, Peterson¹⁴ found that the increase in optical luminosity of the nucleus of this galaxy was accompanied by the appearance of a strong He II 4,686 Å emission line significantly broader than the Balmer lines. This was interpreted by Peterson and Ferland¹ as evidence that the gas producing the He II emission was much closer to the central object and that the increase in luminosity was due to an accretion event of about $0.8 M_\odot$ of material. Figure 1 reproduces the difference (high – low state) spectrum of Peterson¹⁴ and Peterson and Ferland¹. Also shown is the spectrum of the type II supernova SN1983k in NGC4699 at maximum light¹⁵.

The striking similarities between these spectra strongly suggest that Peterson observed the explosion of a type II supernova in the nucleus of NGC5548. Peterson and Ferland did not identify the broad feature next to H γ in their spectrum, but the comparison with SN1983k suggests it could be due to N III emission. Peterson's spectra show that the broad He II line almost disappeared ~ 300 days after the flare, as it did from the spectrum of SN1983k¹⁵. Another important coincidence is that the timescales for the rise and decay of the continuum in NGC5548 are typical of type II supernovae, and of SN1983k in particular. Moreover, the excess luminosity emitted by the 'flare' in the nucleus of NGC5548 in the visual continuum near 5,500 Å is $1.37 \times 10^{15} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ Å}^{-1}$, corresponding to $m_v = 16.1$ or a peak absolute luminosity of -17.4 (assuming Hubble's constant $H_0 = 100 \text{ km s}^{-1} \text{ Mpc}^{-1}$). SN1983k reached a peak absolute luminosity in the V-band of -17.8 .

The only significant difference between the spectra of the 'flare' in NGC5548 and of SN1983k is the slope of the continuum. However, the colour of the 'flare' depends rather critically on the choice of the spectrum representative of the 'low' state of the nucleus and on the observational errors. Peterson¹⁴ finds that after subtraction of the stellar component the nuclear spectra can be fitted with a power-law continuum of slope $\alpha \approx -0.5 \pm 0.1$; nevertheless, if the 'flare' contributes a very small fraction of the total light of the nucleus, the observed uncertainties translate into a much larger uncertainty in the slope of the spectrum of the difference. In the particular case of NGC5548,

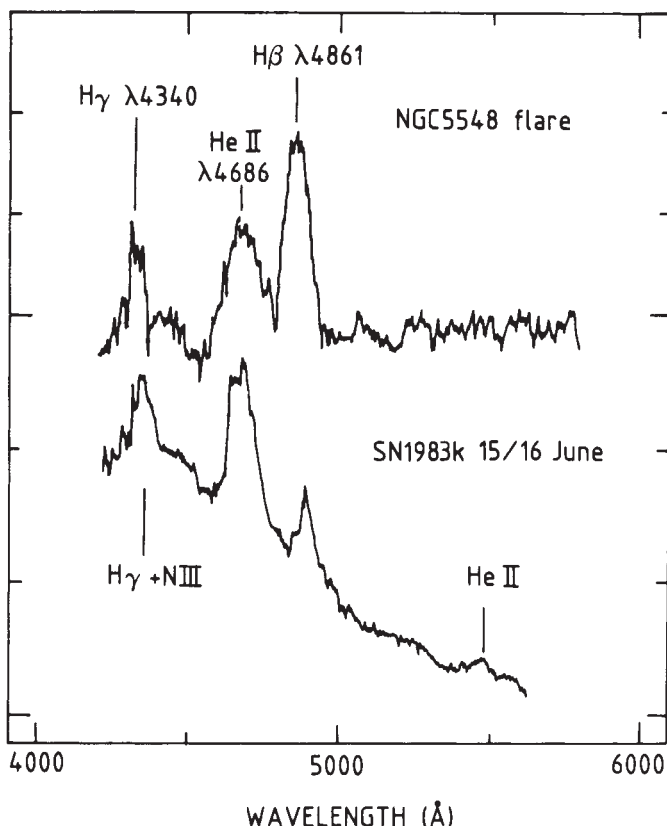


Fig. 1 The top spectrum corresponds to the flare in the optically variable Seyfert type I galaxy NGC5548¹. The bottom spectrum correspond to the type II supernova in NGC4699¹⁵. Fluxes are in relative flux per unit wavelength.

the amplitude of the 'flare' is ~ 0.2 mag, and the uncertainty of ± 0.1 in the slopes of the 'high' and 'low' states translates into an uncertainty of ± 0.7 in the slope of the difference. Therefore, further accurate spectrophotometry is crucial to assess the significance of the apparent discrepancy between the colours of the 'flare' in NGC5548 and of SN1983k.

Supernovae and their remnants are postulated in the starburst model to explain not only the existence of a broad line region and its variability but also to explain the non-thermal radio emission and the X-ray emission from AGNs. The non-thermal radio luminosity of NGC5548 is $3 \times 10^{21} \text{ W Hz}^{-1} \text{ sr}^{-1}$. In the starburst model⁶ this corresponds to a supernova rate of approximately 2 yr^{-1} , in good agreement with the timescale for optical variability observed by Peterson¹⁴. A careful search for nuclear flares should reveal an average of two events per year.

The starburst versus monster controversy continues. The 'accretion event' in NGC5548 may be the first direct observation for a supernova explosion in the nucleus of a Seyfert galaxy. Further spectrophotometric monitoring of AGN variability with adequate temporal coverage and direct searches for nuclear supernovae are clearly very fruitful observational ways to unravel the mystery of the origin of activity in the nuclei of galaxies.

Received 10 December 1987; accepted 29 March 1988.

- Peterson, B. M. & Ferland, G. *Nature* **324**, 345-347 (1986).
- Lynden-Bell, D. *Nature* **223**, 690-694 (1969).
- Rees, M. J. *A. Rev. Astr. Astrophys.* **22**, 471-506 (1984).
- Begelman, M. C. in *Astrophysics of Active Galaxies and Quasistellar Objects* (ed. Miller, J.) 711-752 (University Science Books, Mill Valley, 1985).
- Terlevich, R. & Melnick, J. *Mon. Not. R. astr. Soc.* **213**, 841-856 (1985).
- Terlevich, R., Melnick, J. & Moles, M. in *IAU Symp. 121: Observational Evidences of Activity in Galaxies* (eds Khachikian, E. Ye., Fricke, K. J. & Melnick, J.) 499-519 (Reidel, Dordrecht, 1987).
- Dibai, E. A. & Lyutyi, V. M. *Soviet Astr. Lett.* **2**, 90-92 (1976).
- Lyutyi, V. M. *Soviet Astr.* **23**, 518-523 (1979).

- Ulrich, M. H. *et al. Mon. Not. R. astr. Soc.* **206**, 221-237 (1984).
- Peterson, B. M. & Gaskell, C. M. *Astr. J.* **92**, 552-556 (1986).
- Alloin, D. A. *et al. Astrophys. J.* **308**, 26-35 (1986).
- Lyutyi, V. M. *Soviet Astr.* **21**, 655-664 (1977).
- Dibai, E. A. & Lyutyi, V. M. *Soviet Astr.* **28**, 7-11 (1984).
- Peterson, B. M. *Astrophys. J.* **312**, 79-90 (1987).
- Niemela, V. S., Ruiz, M. T. & Phillips, M. M. *Astrophys. J.* **289**, 52-57 (1985).

Measurements of reflectance spectra of ion-bombarded ice and application to surfaces in the outer Solar System

D. J. O'Shaughnessy, J. W. Boring & R. E. Johnson

Department of Nuclear Engineering and Engineering Physics,
University of Virginia, Charlottesville, Virginia 22901, USA

A number of the icy satellites of the outer planets exhibit interesting hemispherical differences in brightness¹⁻⁵ which have been attributed to enhanced bombardment by the local plasma of one hemisphere. The plasma bombardment is thought to erode the icy surfaces and implant species, thereby altering the surface reflectance spectra^{2,3}, as well as producing fresh plasma⁶. Here we present the first results of laboratory measurements of the wavelength dependence of the alteration of the visible reflectance spectra of H₂O ice irradiated by keV ions. When the implanted species is chemically neutral, absorption is slightly enhanced below $0.55 \mu\text{m}$. For an incident species containing sulphur, a strong absorption feature is produced at $0.4 \mu\text{m}$ corresponding (probably) to S₃. This occurs at too large a wavelength to account for the absorption feature observed at Europa by Voyager and therefore casts doubt on the recent interpretations⁷⁻⁹ of the reflectance data of Europa.

The reflectance spectra of Europa have caused considerable interest in the effect of plasma bombardment. Clark *et al.*¹⁰ deduced that plasma bombardment affected the infrared reflectance and Lane *et al.*¹¹ attributed a band at $\sim 0.28 \mu\text{m}$ on the trailing hemisphere to sulphur implantation from plasma originating at Io. Analysis of the spatial distribution of the absorption seen in the Voyager images taken in the $0.35 \mu\text{m}$ filter also attributed this feature to plasma bombardment⁹. This was accompanied by additional weaker absorption, increasing with decreasing wavelength from $0.6 \mu\text{m}$, the source of which was uncertain⁷⁻⁹. We have therefore measured the reflectance spectra of low-temperature ($\sim 20 \text{ K}$) ice irradiated by ions with energies characteristic of the hot plasma found at Europa by the Voyager spacecraft. The measurements were carried out in a clean environmental system (10^{-9} torr). We first irradiated millimetre-thick samples of H₂O ice with 33 keV Ar⁺ ions, which have a mass and nuclear charge comparable with that of sulphur, but are non-reactive after implantation. We subsequently bombarded fresh samples with SO₂⁺ ions to simulate the effects on the surface of Europa^{7,10}. These ions have average penetration depths of $\sim 20 \text{ nm}$ and tend to split into fast sulphur and oxygen projectiles within the first few monolayers, so that separate sulphur and oxygen atoms are implanted in the ice matrix, simulating the local plasma.

The reflectance spectra ($0.37\text{--}0.65 \mu\text{m}$) were taken with a Bausch and Lomb spectrometer with a resolution of $\leq \pm 0.015 \mu\text{m}$ before and after each irradiation. Typical runs were a few hours. A tungsten lamp was used and the incident light was sent down the beam line, thereby illuminating only the irradiated region. The ratio of the intensity of reflected light at 45° after irradiation to that before irradiation is shown as a function of wavelength in Fig. 1 for three irradiations. The bars indicate the variance in repeated reflectance measurements on the same sample. Typically, ten measurements were made at each data point. The reproducibility with fresh samples is poorer, but the same trend is always observed in any run. The larger error bars at short wavelengths are indicative of the poorer

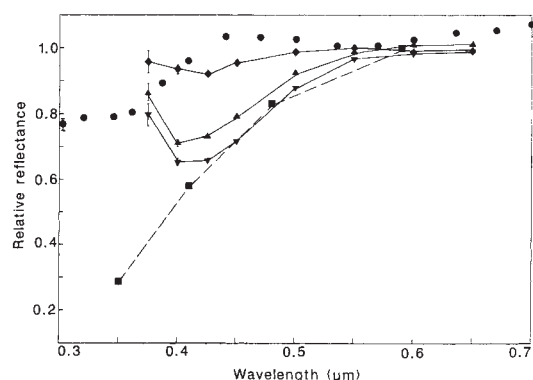


Fig. 1 Relative reflectance against wavelength: solid lines connect the data points for reflectance of irradiated ice sample relative to unirradiated ice ($\blacklozenge$, 33 keV Ar^+ , 3.2×10^{17} ions cm^{-2} ; $\blacktriangle$, 40 keV SO_2^+ , 1.6×10^{17} ions cm^{-2} ; $\blacktriangledown$, 40 keV SO_2^+ , 3.2×10^{17} ions cm^{-2}). Error bars represent root-mean-squared deviation of repeated ratios on each sample. Dashed line connects reflectance of trailing hemisphere of Europa, relative to value at $0.59 \mu\text{m}$ set equal to one, as seen through the filters of Voyager⁷. Circles, ratio of trailing to leading hemisphere reflectance from refs 4 and 19.

response of the spectrometer and the lower levels of light from the source. (For most points, variance is smaller than the symbol used.) In this first set of experiments, the doses used (see Fig. 1) were comparable with ~ 10 – 100 yr of plasma bombardment at Europa at a flux of 10^8 – 10^9 ions $\text{cm}^{-2} \text{s}^{-1}$. Beam current densities were low ($\sim 10 \mu\text{A cm}^{-2}$) to eliminate macroscopic heating (< 1 K rise in samples $1 \mu\text{m}$ thick), so measurements at different currents gave the same result.

In this wavelength region the reflectance spectrum of unirradiated ice is fairly flat, being slightly more absorbing below $\sim 0.45 \mu\text{m}$. After irradiation by Ar^+ , a slightly enhanced absorption is seen at the shorter wavelengths in Fig. 1. As ion bombardment changes the grain size^{10,12}, this effect can be produced by altering grain sizes and roughening the surface. It is also known, however, that plasma bombardment creates O_2 , H_2 and dissociated fragments in the ice, resulting in a more complex chemical material, as well as producing large voids^{13,14}. Based on Fig. 1, a stable icy regolith exposed to a plasma can have a slightly redder spectrum than for H_2O condensed at low temperatures, even in the absence of implanted impurities.

To simulate aspects of the proposed plasma-surface interaction at Europa we also show changes in the reflectance spectrum of ice irradiated by 40 keV SO_2^+ . At longer wavelengths the spectrum shows characteristics similar to that produced in Ar^+ bombardment, but absorbs much more strongly at shorter wavelengths. (As the doses are similar, this suggests that any vacuum contamination does not affect the SO_2^+ results.) This absorption exhibits a band at $\sim 0.4 \mu\text{m}$. Increasing the ion dose deepens this band so that the net reflectance is reduced in the whole region studied here. A band in this region has been identified in the gas phase as S_3 (ref. 15). In this regard, S_4 has an absorption at $\sim 0.53 \mu\text{m}$ (ref. 15) and S_2 absorbs at around $0.32 \mu\text{m}$ (ref. 9) in the gas phase. Reflectance spectra of low-temperature sulphur¹⁶ also exhibit a strong absorption band with onset at $\sim 0.53 \mu\text{m}$ and with the appearance of a possible minimum for large grain sizes at $\sim 0.4 \mu\text{m}$. The absorption observed here, therefore, may be due to the presence of specific sulphur species, such as S_2 or S_3 , in the solid. In contrast, condensed SO_2 frost had a strong absorption below $\sim 0.4 \mu\text{m}$ (ref. 17). Gas-phase SO_2 exhibits a strong absorption band associated with the S–O bond centred at $\sim 0.28 \mu\text{m}$ and extending to $\sim 0.32 \mu\text{m}$ (ref. 11). These experiments therefore indicate that sulphur agglomerates relatively easily under ion agitation of the surface¹⁸.

In Fig. 1, the Voyager reflectance spectrum of the trailing hemisphere of Europa, normalized to unity at $0.59 \mu\text{m}$, is shown

for comparison with our measurements. It is seen that the shape of the absorption above $\sim 0.45 \mu\text{m}$ is very similar to that for implanted SO_2^+ , but the absorption increases with decreasing wavelength down through the $0.35 \mu\text{m}$ filter, a region below the minimum observed here. Also shown here is the ratio of the ground-based measurements of the reflectance on the trailing and leading hemispheres^{4,19}, normalized to unity at $0.59 \mu\text{m}$. This indicates that the principal difference between the hemispherical reflectance arises from an absorption feature below that observed here. Therefore, the reflectance characteristics of the trailing hemisphere of Europa cannot simply be reproduced by the sulphur implantation performed in these experiments, which in many aspects simulates the processes thought to occur at Europa.

We thank B. Hapke, D. Nash and M. McGrath for helpful comments, and T. V. Johnson for a plot of the trailing/leading ratio in Fig. 1. This work was supported by NSF and NASA.

Received 22 January; accepted 29 March 1988.

1. Stebins, J. & Jacobsen, T. S. *Lick Obs. Bull.* **13**, 180–195 (1928).
2. Morrison, D. & Burns, J. A. in *Jupiter* (ed. Gehrels, T.) 991–1034 (University of Arizona Press, Tucson, 1976).
3. Clark, R. N., Fanale, F. P. & Gaffey, M. J. in *Satellites* (eds Burns, J. A. & Matthews, M. S.) Ch. 10 (University of Arizona Press, Tucson, 1986).
4. Johnson, T. V. & Pilcher, C. B. in *Planetary Satellites* (ed. Burns, J. A.) 232–268 (University of Arizona Press, Tucson, 1977).
5. Clark, R. N. & McCord, T. B. *Icarus* **41**, 323–339 (1980).
6. Cheng, A., Haff, P. K., Johnson, R. E. & Lanzerotti, L. J. in *Satellites* (eds Burns, J. A. & Matthews, M. S.) Ch. 9 (University of Arizona Press, Tucson, 1986).
7. Nelson, M. et al. *Icarus* **65**, 129–151 (1986).
8. McEwen, A. *Icarus* **91**, 8077–8097 (1986).
9. Johnson, R. E., Nelson, M. N., McCord, T. A. & Grady, R. *Icarus* (in the press).
10. Clark, R. N., Fanale, F. P. & Zent, A. P. *Icarus* **56**, 233–245 (1983).
11. Lane, A. L., Nelson, R. M. & Matson, D. L. *Nature* **292**, 38–39 (1981).
12. Johnson, R. E. *Icarus* **62**, 344–347 (1985).
13. Johnson, R. E. et al. in *Ices in the Solar System* (eds Klinger, J., Benest, D., Dollfus, A. & Smoluchowski, R.) 301–316 (Reidel, Dordrecht, 1985).
14. Johnson, R. E., Lanzerotti, L. J. & Brown, W. L. *Adv. Space Sci.* **4**, 41–51 (1984).
15. Meyer, B., Stroyer-Hansen, T. & Oommen, T. V. *J. molec. Spec.* **42**, 335–340 (1972).
16. Grady, J. & Veverka, J. *Icarus* **58**, 227–245 (1984).
17. Nash, D. B., Fanale, F. & Nelson, R. *Geophys. Res. Lett.* **7**, 665–668 (1980).
18. Moore, M. H. & Donn, D. B. *Icarus* (in the press).
19. Johnson, T. V. *Icarus* **14**, 94–111 (1971).

Field evidence for the oxidation of SO_2 by H_2O_2 in cap clouds

G. P. Gervat*, P. A. Clark*, A. R. W. Marsh*,
I. Teasdale*, A. S. Chandler†, T. W. Choularton†,
M. J. Gay†, M. K. Hill†‡ & T. A. Hill†

* Central Electricity Generating Board, Central Electricity Research Laboratories, Kelvin Avenue, Leatherhead, Surrey, UK

† Atmospheric Physics Research Group, UMIST, Manchester, UK

The reaction of SO_2 , dissolved in cloud droplets, with atmospheric oxidants such as H_2O_2 and ozone is thought to be an important pathway for the production of atmospheric sulphate, and may be significant in determining the pattern of acidic deposition over a region. Direct observation of in-cloud oxidation of SO_2 by these species, however, has proved extremely difficult. This is especially true for H_2O_2 because the reaction is so rapid. We have used a new experimental design to investigate the dominant oxidation pathways of SO_2 in the conditions that occur in natural clouds. This involved modifying SO_2 levels in capping cloud on a peak in the northern Pennine hills of England by releasing SO_2 locally upwind and then measuring changes in cloud and air composition. Here we present initial results from an experiment performed on 12 December 1985. Under the conditions encountered it was found that sulphate production by both ozone and H_2O_2 was negligible. The destruction of H_2O_2 by reaction with SO_2 , however, was

‡ Present address: HOLTECH, 4 Riggside, Harwood-in-Teasdale, Barnard Castle, Co. Durham, UK.

clearly demonstrated. By comparing H_2O_2 in the presence and absence of released SO_2 , the rate of destruction was estimated and found to be comparable with that expected from laboratory measurements of the reaction rate constant, and our knowledge of the cloud physics.

Controlled emissions of both SO_2 and SF_6 (as a reference tracer) were made in a molar ratio of 120:1 from the top of a 20 m tower ~5 km upwind of Great Dun Fell (847 m above sea level, 54°41' N, 2°27' W). A mobile laboratory continuously travelled along a mine road, located approximately 170 m below the summit, measuring plume SO_2 and SF_6 concentrations (see Fig. 1). Cloud base was at an altitude corresponding to a distance of ~1–2 km upwind of the summit. Measurements of cloud and air composition were made both at the hill summit and at a van equipped with instruments at the western end of the mine road. Figure 2a shows the Na^+ , NH_4^+ , NO_3^- , Cl^- and SO_4^{2-} concentrations in cloud-water at the summit together with cloud liquid-water content. The upward drift in ionic concentrations is primarily a result of changes in liquid water content. Figure 2a also demonstrates that no SO_4^{2-} above background levels can be detected from the measurements on bulk samples and hence no significant sulphate production mechanism was operating over the timescale of a few hundred seconds. This is confirmed by Fig. 2c which shows that the total SO_4^{2-} in upwind aerosol was comparable with the cloudwater concentration (expressed as equivalent aerosol).

Concentrations of ozone were $\sim 22 \pm 3$ p.p.b.v. (parts per 10^9 by volume) and no change resulting from the SO_2 release could be detected above natural variability and instrument noise. Indications of some SO_2 oxidation by H_2O_2 , albeit at a very low level, can, however, be seen in the measurements of H_2O_2 in cloud water (Fig. 2b). Background levels (that is, outside the plume) at the summit were about $0.6 \mu\text{M}$.

To eliminate changes in concentration resulting from the natural variability of cloud liquid-water content, a total equivalent concentration may be defined. This is the total amount of a species per unit volume of air and is derived by conceptually evaporating all the cloud droplets in that air parcel while assuming that the aqueous species had been in equilibrium with the gas phase. For H_2O_2 , this is given by:

$$\text{H}_2\text{O}_{2(\text{tot})} = RT(v+1/H)[\text{H}_2\text{O}_{2(\text{aq})}]/p$$

where $\text{H}_2\text{O}_{2(\text{tot})}$ is the volume mixing ratio of H_2O_2 , R is the universal gas constant ($\text{J K}^{-1} \text{mol}^{-1}$), T is the temperature (K), v is the volume mixing ratio of cloud water, $[\text{H}_2\text{O}_2]$ is the H_2O_2 concentration (mol l^{-1}), H is the Henry's law constant $[\text{H}_2\text{O}_{2(\text{aq})}]/[\text{H}_2\text{O}_{2(\text{g})}]$, and p is the ambient pressure (N m^{-2}). The total background equivalent concentration of H_2O_2 was 8 ± 1 p.p.t.v. (parts per 10^{12} by volume) at the summit (Fig. 2d) and 7 ± 1 p.p.t.v. at the van site. Despite this low concentration, the effects of the SO_2 release between 12:42 and 14:00 hours were clearly resolvable as depletion of the background H_2O_2 concentration. Several dips in H_2O_2 concentration may be identified as parts of the release plume were advected over the summit.

These measurements represent an unequivocal demonstration of the reaction of SO_2 and H_2O_2 in cloud. They may be interpreted in a simple way by considering parcels of air as they travel from cloud base to the summit. Because SO_2 was in considerable excess over H_2O_2 it can be shown that:

$$\ln([\text{H}_2\text{O}_2]_o/[\text{H}_2\text{O}_2]_i) \propto [\text{SO}_2]$$

where $[\text{SO}_2]$ is the total equivalent concentration of SO_2 , $[\text{H}_2\text{O}_2]$ is the total equivalent concentration of H_2O_2 and 'i' and 'o' refer to inside and outside the plume at the hill summit. The constant of proportionality depends on the reaction rate-constant, transit time and the history (temperature, liquid-water content and so on) as it moves up the hill. Application of this equation to the observations at the summit shows that different parts of the SO_2 plume exhibit similar pseudo first-order behaviour (see Fig. 3),

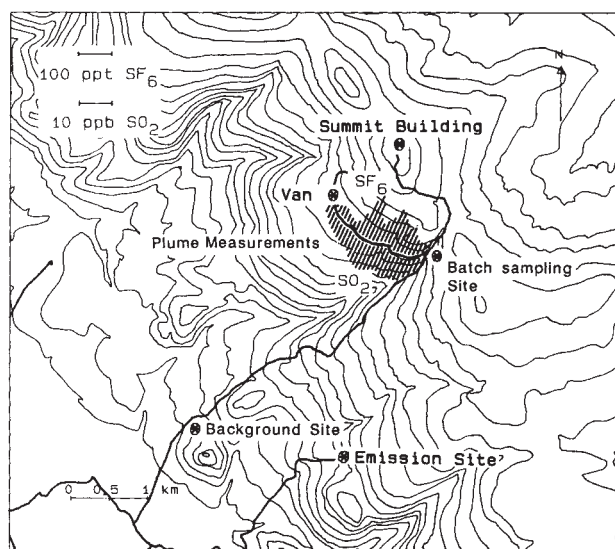


Fig. 1 Map of Great Dun Fell showing the gas release and measurement sites. Plume SO_2 and SF_6 concentrations measured during a single traverse of the plume along the mine road by the mobile laboratory, between 14:05 and 14:16 hours, are also shown.

despite different degrees of dilution and hence different mixing histories.

To estimate an overall field reaction rate-constant, the cap-cloud model of Hill, Choularton and Penkett¹ that takes into account the varying conditions encountered by parcels of air moving up the hill, was used. The concentration of SO_2 in an air parcel entering cloud base was set equal to one of the peak concentrations observed at the summit (see Fig. 2d). The gas phase H_2O_2 concentration entering cloud base was assumed to be equal to the total equivalent concentration of liquid phase H_2O_2 measured outside the plume. Then the evolution of the H_2O_2 concentrations in solution and in the gas phase, as the cloud droplets grew over the 160-transit from cloud base to the hill top, was computed. The liquid-water content was such that it agreed with the measurements at the two sites. The reaction rate-constant in the model was adjusted and the model re-run until the predicted concentration of H_2O_2 at the summit was equal to that observed.

The primary error in this procedure arises because in reality the SO_2 plume mixes continuously with the surrounding cloud as it ascends the hill. This has two effects: the SO_2 concentrations may be higher at cloud base than at the summit; and, more importantly, extra H_2O_2 can be incorporated into the SO_2 plume. Furthermore, Fig. 2d shows that SO_2 concentrations vary considerably at the summit and therefore that this mixing process is not uniform. Analysis of Fig. 3 suggests that these effects are of secondary importance because different parts of the plume exhibit similar reaction rate-constants. This has been confirmed by re-running the model in two ways. In the first case, the ascending plume was allowed to entrain air and liquid water continuously from the surrounding cloud with the assumption that the mixing occurred homogeneously across the plume. In the second, entrainment was treated as a set of discrete events to reproduce the observed plume inhomogeneities. Both H_2O_2 and SO_2 were assumed to be in equilibrium with the liquid phase in the air parcels, at all times. Using these methods, a crude description of an expanding plume was achieved which showed that neglecting the effects of mixing into the plume leads to a 10–30% underestimate of the $\text{SO}_2/\text{H}_2\text{O}_2$ reaction rate.

By running the model for those concentrations used to obtain the points in Fig. 3, a reaction rate-constant

$$-\frac{d}{dt}[\text{H}_2\text{O}_{2(\text{aq})}] \times \frac{1}{[\text{H}_2\text{O}_{2(\text{aq})}][\text{S}_{\text{iv}(\text{aq})}]} = 200 \pm 50 \text{ M}^{-1} \text{ s}^{-1}$$

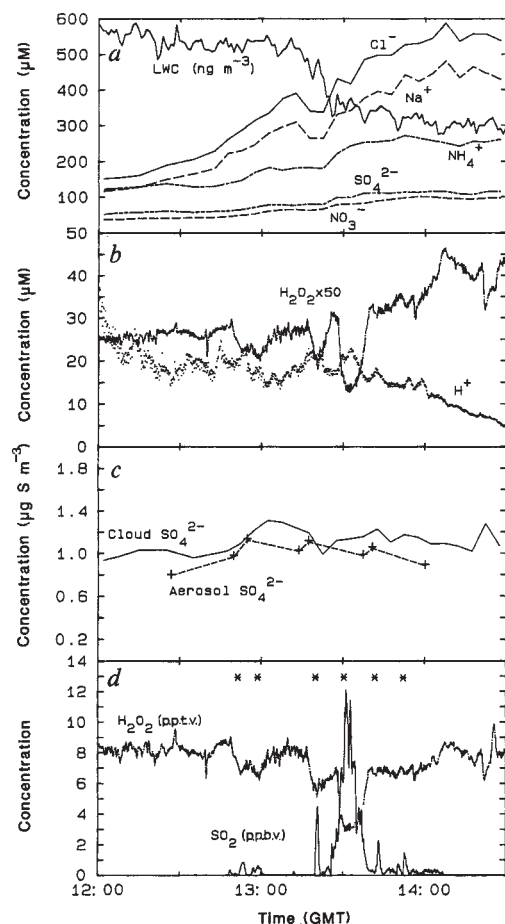


Fig. 2 Results from the summit site on 12 December 1985. *a*, Cloud water composition and cloud liquid-water content (LWC); ionic species are derived from bulk samples of ~5 min duration. *b*, H^+ derived from continuous pH measurements and aqueous H_2O_2 . *c*, Comparison of SO_4^{2-} in cloud, expressed as equivalent aerosol, and upwind measurements from filters with a sampling time of ~20 mins. *d*, Total equivalent concentration of SO_2 (derived from SF_6) and H_2O_2 . Significant H_2O_2 dips associated with released SF_6 and SO_2 , used to derive Fig. 3, are shown by *.

(at approximately 278 K and pH 4.75) for the H_2O_2/SO_2 reaction was derived. The estimated random error in the value of the constant is attributable to the effects of mixing and to uncertainties in the measurements. The latter were approximately $\pm 10\%$ for SO_2 , $\pm 10\%$ for liquid-water content and $\pm 5\%$ for H_2O_2 . Data from a single observation of the passage of the SO_2 plume at the mine road provided an estimate of $600 \pm 200 M^{-1} s^{-1}$ for the rate constant. At the same temperature and pH, recent laboratory determinations of the rate constant by different authors range from 349 to $640 M^{-1} s^{-1}$ (refs 2–5). The field-derived results encompass the laboratory values and we can therefore be confident that on this occasion oxidation of SO_2 in cloud by H_2O_2 was proceeding in a manner predictable from laboratory measurements.

The experiments at Great Dun Fell are the first successful attempt at changing cloud chemistry by artificially adjusting the input SO_2 concentration. Results have been obtained which demonstrate the reaction of SO_2 and H_2O_2 in cloud and allow an effective rate constant to be estimated which is not very different from values measured in the laboratory. During similar experiments in springtime, when photochemical activity was higher, significant sulphate production was observed as a result of released SO_2 being oxidized by both H_2O_2 and O_3 (manuscript in preparation). The principle of introducing controlled

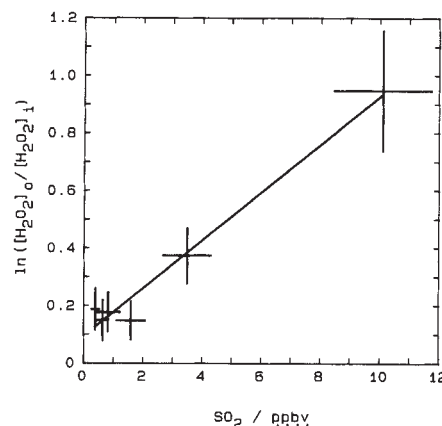


Fig. 3 Plot of $\ln([H_2O_2]_o/[H_2O_2]_i)$ against SO_2 at the summit on 12 December 1985. Error bars are estimated from calibration uncertainty and signal noise level, and represent approximately one standard deviation either side of the mean.

emissions upwind of cloud is now firmly established and the technique will be used for other species and such as NH_3 , NO_x , HCHO.

This work is published by permission of the Central Electricity Generating Board.

Received 18 November 1987; accepted 19 February 1988.

- Hill, T. A., Choulaton, T. W. & Penkett, S. A. *Atmos. Envir.* **20**, 1763–1771 (1986).
- McElroy, W. J. *Atmos. Envir.* **20**, 427–438 (1986).
- Kunen, S. M., Lazrus, A. L., Kok, G. L. & Heikes, B. G. *J. geophys. Res.* **88**, 3671–3674 (1983).
- McArdle, J. V. & Hoffman, M. R. *J. Am. chem. Soc.* **87**, 5425–5429 (1983).
- Martin, L. R. & Damschen, D. E. *Atmos. Envir.* **15**, 1615–1621 (1981).

Ice-age hurricanes and tropical storms

Jay S. Hobgood* & Randall S. Cerveny†

* Atmospheric Sciences, Department of Geography, Ohio State University, Columbus, Ohio 43210–1361, USA

† Department of Geography, Arizona State University, Tempe, Arizona 85287-0104, USA

We compare Pleistocene tropical cyclones with present-day storms in this study through the use of a three-dimensional numerical model simulating hurricane development and processes¹. The characteristics of Atlantic ice-age tropical storms are derived from reconstructions of the 18,000 BP climate (the Last Glacial Maximum or LGM) and the GISS (Goddard Institute of Space Studies) General Circulation Model's LGM simulation^{2,3}. The simulation results indicate that although the structure of the ice-age atmosphere was colder and more stable, it had the potential to support development of tropical cyclones. Such storms, however, would be substantially weaker than contemporary storms. This investigation demonstrates the value of integrating numerical models of varying spatial scales in climatic research.

Although palaeoclimatologists have not attempted reconstruction of individual Pleistocene tropical storms, it is generally agreed that the tropics were much colder and drier than present-day^{4–11}. A hurricane-frequency study¹² derived an empirical relationship between sea-surface temperature (SST) and the frequency of tropical storms that implied tropical storm frequencies during glacial episodes were much lower than present. Many studies based on deep-sea cores analysis^{13–18} have used this relationship with reconstructed values of SSTs to suggest decreases in hurricane frequency during the LGM.

The 1976 CLIMAP reconstruction¹⁷ indicated the LGM area

of 26 °C water is reduced to a third of its present size. A later CLIMAP reconstruction¹⁸ suggested that large areas of the LGM tropical and subtropical oceans had SSTs as warm as, or warmer than, today, although its reconstruction of the tropical Atlantic was consistent with the earlier study. This reconstruction gave temperatures 2 °C colder than present conditions for the tropical Atlantic/Caribbean. Detailed studies of the accuracy and precision of these estimates have placed the uncertainty in temperature estimates at 1 °C for tropical latitudes. Our study incorporates these conditions, together with conditions derived from the GISS General Circulation Model (GCM), into a fine-resolution numerical tropical cyclone model.

The basic components of the hurricane model used in this study have been discussed earlier¹. The model is based on the primitive equations of motion expressed in the sigma-coordinate system¹⁹. Four vertical layers represent the boundary layer and the lower, middle and upper tropospheric layers. The horizontal grid uses a uniform spacing of 25 km. Horizontal velocity components are calculated on a grid offset by 45° to reduce truncation error, arising during calculation of the pressure gradient force, with minimal computational expense²⁰. Time integration is performed using Matsuno's method²¹ and the time step is 37.5 s.

The thermodynamic effects of convection are determined from a modification of the Kuo parameterization²². In this scheme, water available for convection is equal to the horizontal convergence of moisture in the boundary layer and evaporation from the ocean, or

$$M_t = -\nabla \cdot p^* \bar{V} q_4 \delta\sigma_4 + Q_{SEA} \quad (1)$$

where M_t is the moisture supply available for convection, p^* is surface pressure, $\bar{V}$ is horizontal velocity, q_4 is specific humidity in the boundary layer, $\delta\sigma_4$ is the sigma thickness of the boundary layer, and Q is evaporation from the ocean. In this parameterization²², percentages of the moisture available for convection assigned to the moisture enrichment in an atmospheric column (b) and to the latent heating ($1-b$) are specified explicitly. The equation for the calculation of b is given by

$$b = \left[\frac{1 - \sum_{k=1}^4 \frac{RH_k \delta\sigma_k}{1 - RH_{cr}} \right]^n \quad (2)$$

where RH_k is relative humidity at level k , $\delta\sigma_k$ is sigma thickness of layer k , RH_{cr} is critical relative humidity ($RH_{cr} = 0.5$) and n is an empirically determined exponent ($n = 1.0$) (ref. 23).

The vertical distribution of latent heating²² is given by

$$p^* \frac{\partial T}{\partial t_c} = \frac{L(1-b)M_t(T_c - T)}{c_p \sum_{k=1}^4 (T_c - T) \delta\sigma_k} \quad (3)$$

where T_c is cloud temperature, T is environmental temperature, L is latent heat of evaporation and c_p is the specific heat at constant pressure of dry air. The vertical transfer of moisture by convection²² is specified by

$$p^* \frac{\partial q}{\partial t_c} = \frac{bgM_t(q_c - q)}{\sum_{k=1}^4 (q_c - q) \delta\sigma_k} \quad (4)$$

where q_c is specific humidity inside the cloud, q is specific humidity of the environment and g is gravitational acceleration.

The second version of this study's model explicitly includes, in the thermodynamic equation, radiative fluxes of energy¹. These radiative values are computed in part using formulae developed for calculating daily insolation values associated with variations in the Earth's orbital elements (obliquity, eccentricity and equinoctial precession)^{24,25}. Orbital parameters obtained in this study were comparable with results from other models²⁵. Evidence indicates that past variations in terrestrial orbital parameters have played major roles in development of glacial-interglacial episodes²⁶. Recent studies suggest that radiative fluctuations (particularly the diurnal cycle) may strongly influence the size and intensity of contemporary tropical storms^{1,27}.

Pleistocene surface boundary conditions for 20° N over the

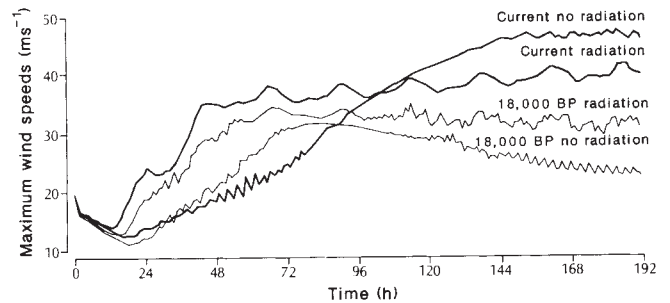


Fig. 1 Maximum wind speeds (m s^{-1}) for radiative and non-radiative simulations under present-day and 18,000 BP conditions.

tropical Atlantic for July were adapted from palaeoclimatic data and simulations^{2,18,25}. Initial atmospheric conditions (namely, temperature and specific humidity profiles) for the 18,000 BP tropics were derived from global LGM July simulations by the GISS GCM (refs 2 and 3), which are comparable with other model LGM simulations²⁵. Horizontally weighted averages of these parameters for the North Atlantic area (centred at 22° N, 70° W) were derived using the GISS fine resolution GCM's 4° latitude by 5° longitude grid locations. The initial vertical profiles of specific humidity and temperature used in the present-day control simulations are derived from empirical data²⁸. In all simulations, these conditions are initialized as horizontally constant.

A circularly symmetrical surface pressure pattern, representing the 'seedling' tropical depression, and the gradient wind equation initially generate the weak initial vortex at the centre of the model. After initialization, the finite difference forms of the basic primitive equations are integrated forward in time to simulate the growth and development of the tropical cyclone. Using current climatic conditions for the tropical Atlantic, both radiative and non-radiative versions of the model reproduce features common to tropical cyclones (that is, strong low-level inflow, upper level divergence, a warm core in middle and upper troposphere, and spiral rain bands in the mature hurricane) and are consistent with other models and observational data¹.

Four simulations were performed to compare the relative characteristics of present-day tropical cyclones with those of 18,000 BP. Maximum wind speeds from these simulations indicate that the thermodynamic state of the atmosphere over the LGM Atlantic was less capable of supporting tropical cyclone development than is the present atmosphere (Fig. 1). When the radiative effects are excluded from the Pleistocene simulation, the storm reaches maximum wind speeds of 32 m s^{-1} at 85 h (slightly below the minimal hurricane wind speed of 33.5 m s^{-1}), and decays slowly thereafter. By contrast, the present-day non-radiative tropical cyclone reaches maximum wind speeds of 47 m s^{-1} and is able to maintain its intensity throughout the 8-day simulation. The slight variation in maximum winds evident towards the end of all simulations is partially due to truncation errors in the finite difference equations. But as the magnitude of the difference between simulations is much greater than this small variation, comparisons in storm development and growth between simulations are reflective of changes in climate and radiative conditions.

When radiative fluxes are included in the simulations, the maximum wind speeds again indicate that the LGM storm is weaker than the present-day storm, but the difference is not as pronounced as the non-radiative simulation (Fig. 1). The contemporary storm reaches maximum wind speeds of 42 m s^{-1} , whereas the LGM storm sustains minimal hurricane wind speeds of 35 m s^{-1} . In addition, with the inclusion of the radiative fluxes, the Pleistocene storm does not exhibit the pronounced decrease in intensity that was evident without these fluxes. As with the non-radiative simulations, the present-day simulation develops and maintains lower minimum sealevel pressures.

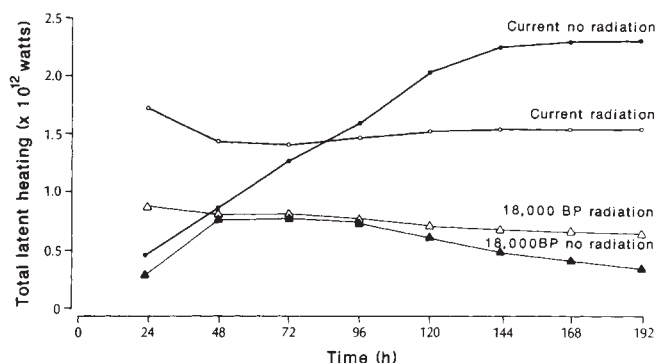


Fig. 2 Total latent heating for radiative and non-radiative simulations under present-day and 18,000 BP conditions.

During the first stages of the simulations, inclusion of radiative fluxes produces more intense storms under both ice-age and contemporary conditions. The controlling mechanism during this developmental stage is nocturnal cooling of the cloud tops, which aids in destabilizing the atmosphere and increasing convection. Increased convection releases greater amounts of latent energy to the upper reaches of the storm and produces a more intense system. This effect is particularly important for the LGM, during which the atmosphere over the tropical Atlantic was markedly more stable than present-day and the influx of solar radiation for the tropics was similar to present-day. By the later stages of simulation, the radiative fluxes produce a weaker present-day tropical cyclone and, interestingly, a more intense Pleistocene storm than when these fluxes are not explicitly included.

Although this appears initially contradictory, it may be reconciled by examining the differences between contemporary and ice-age atmospheres. The warm troposphere of the present-day emits greater radiation to space than does its colder Pleistocene counterpart. As latent energy is released by the intensifying contemporary storm, the middle and high tropospheric temperatures increase. By the later stages of the simulation, the amount of radiation emitted to space has progressively increased, thus resulting in a less intense tropical cyclone than one in which radiative fluxes are excluded. Conversely, the colder temperatures result in less radiative energy loss by the whole system for 18,000 BP. This allows the previously mentioned destabilizing impact at the cirrus canopy top to generate a slightly more intense system than the storm produced when radiative fluxes are included.

Support for this conclusion is visible in the magnitudes of total latent heat energy released by the four simulated tropical cyclones (Fig. 2). The release of latent energy generates the warm upper core and intensifies the tropical cyclones. Under present conditions, some of this energy is used to offset the radiative loss to space and results in a slightly weaker storm by the later stages of the simulation. Although this occurs to a lesser extent at 18,000 BP, the more important effect of destabilization of the atmosphere produces slightly more convection and releases more latent energy. The result is a tropical storm more intense than would occur without the radiative fluxes. But both ice-age storms show less latent energy release and convection, resulting in smaller precipitation totals over the storm's lifetime and, consequently, to drier conditions in those areas affected by Atlantic tropical storms. This supports findings of an arid environment in Florida during the LGM (refs 10, 11).

Examination of other variables (such as pressures, precipitation rates, vertical velocities and relative humidities) show similar patterns to those of wind speed and latent heat release. By all measures of intensity, simulations of contemporary tropical cyclones are more intense than those which develop for LGM conditions, although, with radiative fluxes, the differences are significantly lessened. Our results demonstrate that, whereas

Pleistocene systems were substantially weaker than those which are produced today, the ice-age tropical atmosphere of the North Atlantic could have thermodynamically supported the development of minimal hurricanes.

We sincerely thank Drs Hansen, Rind and Peteet for the results of their GCM simulation of 18,000 BP conditions, as well as their comments regarding the CLIMAP boundary conditions. We also appreciate data and comments furnished by Drs Kutzbach and Gates regarding LGM simulations by the NCAR and OSU climate models.

Received 19 November 1987; accepted 29 March 1988.

- Hobgood, J. S. *J. Atmos. Sci.* **43**, 2901-2922 (1986).
- Hansen, J. *et al.* in *Climate Processes and Climate Sensitivity* (eds Hansen, J. E. & Takahashi, T.) 130-163 (American Geophysical Union, Washington, 1984).
- Hansen, J. *et al. Mon. Weath. Rev.* **111**, 609-662 (1983).
- Van der Hammen, T., Barends, J., De Jong, H. & De Veer, A. A. *Paleogeogr. Paleoclimatol. Paleocool.* **32**, 247-340 (1981).
- Felder, W., Hayden, B. & Dolan, R. *J. Geol.* **83**, 455-461 (1979).
- Imbrie, J. & Kipp, N. G. *The Late Cenozoic Glacial Ages* (ed. Turekian, K. K.) (Yale University Press, New Haven, 1971).
- Marsaglia, K. M. & Klein, G. D. *J. Geol.* **91**, 117-142 (1983).
- Alford, J. J. & Holmes, J. C. *Ann. Assoc. Amer. Geog.* **73**, 395-403 (1985).
- Bonatti, E. & Gartner, S. *Nature* **224**, 563-565 (1973).
- Wright, H. E. *Q. Res.* **15**, 113-125 (1981).
- Watts, W. A. *Geology* **3**, 344-346 (1975).
- Wendland, W. M. *J. appl. Met.* **16**, 477-481 (1977).
- Shackleton, N. J., Hall, M. A., Line, J. & Shuxi, C. *Nature* **306**, 319-322 (1983).
- Dansgaard, W. & Tauber, H. *Science* **166**, 499-502 (1969).
- Lynts, G. W. & Judd, J. B. *Science* **171**, 1153-1144 (1971).
- Emiliani, C. *Science* **178**, 393-401 (1972).
- CLIMAP project members *Science* **191**, 1131-1137 (1976).
- CLIMAP project members *Seasonal Reconstruction of the Earth's Surface at the Last Glacial Maximum* (Geological Society of America, Map & Chart Series MC-36, 1981).
- Phillips, N. A. *J. Met.* **14**, 184-185 (1957).
- Anthes, R. A. *Mon. Weath. Rev.* **100**, 461-476 (1972).
- Matsumoto, T. *J. Met. Soc. Japan* **44**, 76-84 (1966).
- Kuo, H. L. *J. Atmos. Sci.* **22**, 40-63 (1974).
- Anthes, R. A. *Mon. Weath. Rev.* **105**, 270-286 (1977).
- Berger, A. L. *J. Atmos. Sci.* **35**, 2362-2367 (1978).
- Kutzbach, J. E. & Guetter, P. J. *J. Atmos. Sci.* **43**, 1726-1759 (1986).
- Berger, A. L., Imbrie, J., Hays, J., Kukla, G. & Saltzman, B. *Milankovitch and Climate* (Reidel, Dordrecht, 1984).
- Browner, S. P., Woodley, W. L. & Griffith, C. G. *Mon. Weath. Rev.* **105**, 856-864 (1977).
- Herbert, P. J. & Jordan, B. R. *Mon. Weath. Rev.* **87**, 217-221 (1959).

Clues to Arctic soil erosion from cryo-electron microscopy of smectite

Kazuo Tazaki*, W. S. Fyfe* & M. Iwatsuki†

* Department of Geology, The University of Western Ontario, London, Ontario, N6A 5B7, Canada

† JEOL Ltd, 1418 Nakagami, Akishima, Tokyo 196, Japan

Soil erosion is attracting increasing attention worldwide as demands on land use increase with the growing human population. Arctic soils appear to be extremely sensitive to erosion by wind and water. Because of ice sheet erosion and massive exposure of fresh rock Arctic soils are often young and dominated by swelling clays. Freeze-thaw cycles in such water-adsorbing materials must influence the mechanics of the soil horizons. In this study using high-resolution cryo-electron microscopy we show the formation of ice in such clay structures and show that the freezing process can deform the clay structure and even produce particles of aerosol dimensions.

The Worldwatch Institute¹ has estimated that present global loss of topsoil is about 0.7% per year and has termed the problem "the quiet crisis in the world economy"²⁻³. In Canada and the northern United States, increasing attention is being given to the problem of loss of topsoil by wind and water erosion, problems which have been exacerbated by the widespread use of herbicides.

We have been intrigued by the small quantities of soil present in the northern latitudes of Canada and formed since retreat of the great ice sheets. It is not infrequent to find basaltic rocks

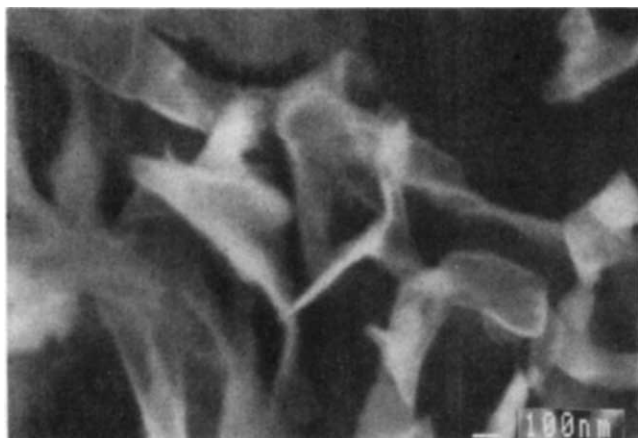


Fig. 1 Cryo-scanning electron micrograph of freeze-dried Na-smectite. The micrograph shows high-porosity and almost single-plate card-house structure with edge-to-face contacts.

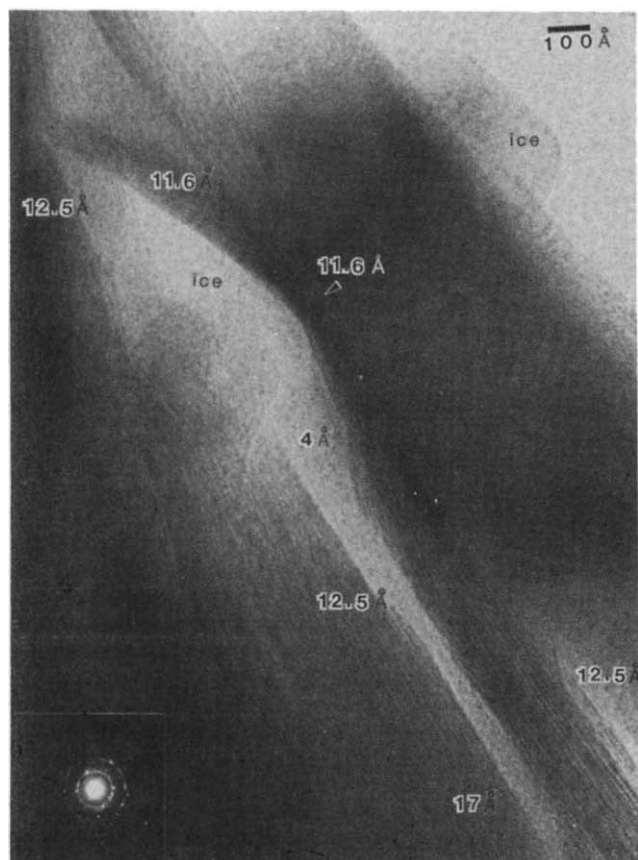


Fig. 2 Cryo-transmission high-resolution electron micrograph of frozen-hydrated Na-smectite and the selected-area electron diffraction pattern showing the deformation of the basal spacing (arrow) and the shrinkage to 11.6 Å from 12.5 Å and 17 Å around an ice crystal. The ice crystal shows 4 Å lattice images. The electron diffraction shows the spacings of both ice and smectite as 5.82 Å (003*), 3.88 Å (ice), 3.05 Å (006*), 2.25 Å (ice and 220*, 040*), 2.06 Å (009*), 1.50 Å (330*, 060*), and 1.28 Å (ice and 260*, 400*); * *hkl* for smectite.

with only trivial films of weathering products, a situation which can be compared with soil formation on the young basalts of Hawaii. Admittedly there can be differences in the primary mineralogy, for most of the older basaltic rocks have greenschist facies mineralogy, but this alone does not seem to explain the very large differences. The vast loess deposits of Europe, Asia⁴

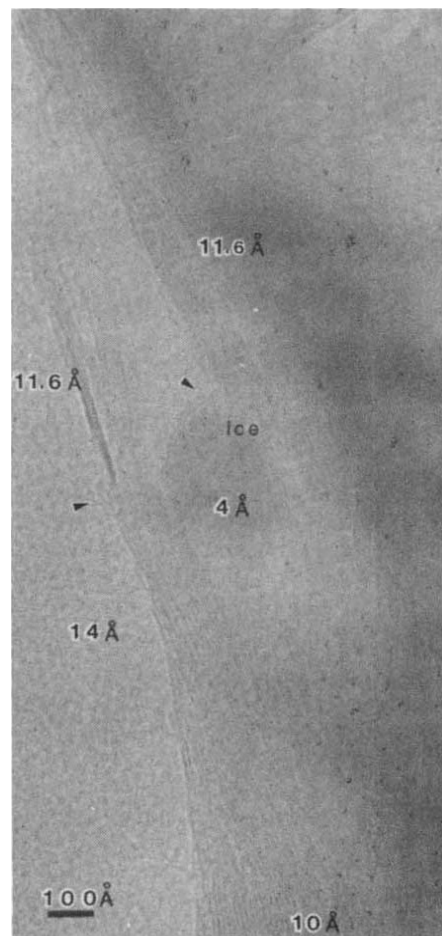


Fig. 3 Cryo-transmission high-resolution electron micrograph of frozen-hydrated Na-smectite showing the damage of the basal spacing near ice crystal (arrows). The basal spacing has changed to 11.6 Å and 10 Å from 14 Å with dislocation by freezing. Parts of the material near ice are ruptured to small fragments (20 Å) or even appear to be amorphous.

and North America⁵ indicate the scale and transport directions of the chief Arctic wind systems.

A characteristic feature of the more northern latitudes is the long period of frozen conditions coupled to zones of high winds. For example, at Resolute in the North West Territories average wind speeds are 35.5 km h⁻¹. Such observations led us to question whether or not there could be a relation between the freeze-thaw cycle and surface wind erosion. It should be noted that in many northern situations oscillations through the freeze-thaw cycle will occur many times in a single year.

When fresh rocks are weathered, the dominant clay types to form are the smectites that have the ability to absorb extra layers of water in the structure. We wished to examine what happens when such clays were taken to very low temperatures.

Using a standard Na-smectite (from the Black Hills, Wyoming), changes in the Na- and Mg-saturated types were studied by X-ray diffraction at liquid nitrogen temperatures. Typical changes in clay spacings and the formation of ice in the clay were 20–22% shrinkage of the basal spacing (001) of smectite at -35 °C. The basal spacing of 15.2 Å of Na-smectite was decreased to 12.2 Å by freezing. The Mg-saturated smectite also showed the decrease in basal spacing from 20.3 Å to 15.9 Å by freezing. For electron microscopy small droplets of smectite suspended in deionized distilled water were mounted on a cold Cu-stub, and then instantly the sample holder was placed in contact with a cold Cu-brick cooled by liquid nitrogen, and finally the sample holder was dipped in liquid nitrogen. After

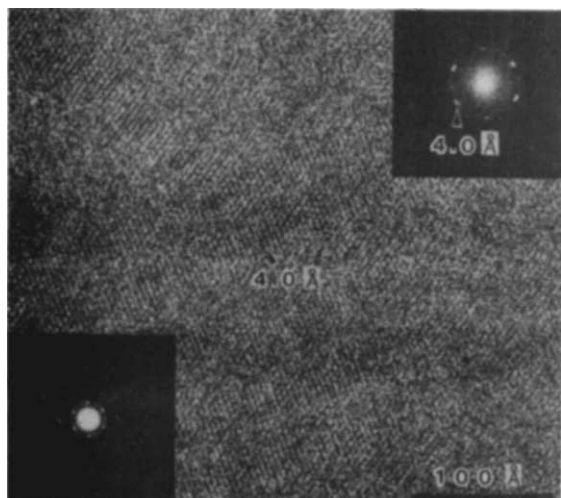


Fig. 4 Cryo-transmission high-resolution electron micrograph of the ice showing 4.0 Å lattice images with the selected-area electron diffraction pattern and the optical diffraction pattern.

removing surface frost and ice by evaporation, particles were coated with gold. A super-conducting lens was used in most measurements.

Of great interest are results from cryo-SEM and cryo-TEM, using a JEOL JSM-T 330A/CRU at liquid nitrogen temperatures and a JEOL TEM-2000 FX with a liquid helium stage. Descriptions of the techniques for use of cryo-SEM with biological samples⁶⁻¹⁰, plant materials¹¹ and soils^{12,13} and cryo-TEM techniques for viruses¹⁴, AgBr and poly-*p*-xylylene¹⁵ are well documented. Iwatsuki *et al.*¹⁶ have studied the advantage of superconducting lens techniques which maintain high-resolution capability for specimens sensitive to radiation damage. Selected-area electron diffraction patterns were calibrated with an evaporated gold film as standard. A <0.2 µm size fraction cryo-SEM of air-dried smectite showed the formation of close-packed aggregates of particles with low porosity. In contrast, freezing and freeze-drying of water suspensions produced the 'house of cards' structure (or aerogel) texture shown in Fig. 1. Cavities represent sites previously occupied by ice. The density of such material must be extremely low and the microstructure a perfect wind catcher. The pore size developed depended on the actual concentration of the suspension, the clay/water ratio. In cryo-TEM (at liquid He temperatures) details of the freezing process were revealed. As shown in Figs 2 and 3, ice particles 100–200 Å in size form in the smectite interlayers. The ice formation clearly deforms or even ruptures the silicate layer structure. Perfect ice (form *Ih*) lattice images and spacings were obtained, but a domain structure with partially amorphous regions is apparent (Fig. 4).

Clearly, rapid freezing can produce a highly porous material *in situ* or in the laboratory¹⁷ and can cause a fundamental reduction in particle size. It appears that extremely fine particles of typical aerosol size can be produced by the freeze-thaw cycle.

Received 25 January; accepted 28 March 1988.

1. Brown, L. R. & Wolf, E. C. *Worldwatch Pap.* No. 60 (1984).
2. *A Growing Concern: Soil Degradation in Canada* (Science Council of Canada, 1986).
3. Sparrow, H. O. *Soil at Risk* (The Senate of Canada, Ottawa, Ontario, 1984).
4. Chu, S. M. *Land Resources in the Loess Plateau of China* Vol. 2 (Shaanxi Science and Technique, Shaanxi, 1986).
5. Humphries, D. W. in *The Encyclopedia of Sedimentology* (eds Fairbridge, R. W. & Bourgeois, J.) 284–288 (Dowden, Hutchinson and Ross, Pennsylvania, 1978).
6. Suzuki, T. & Kamaya, A. *JEOL Scanning Microscope* 26, 2–5 (1986).
7. Inoue, T., Shimakura, S., Sato, T. & Koike, H. *J. Scanning Electron Microscopy (Japan)* 11, 26–28 (1982).
8. Shimakura, S. & Inoue, T. *J. Electron Microsc. (Japan)* 19, 172–176 (1985).
9. Inoue, T. in *Science of Biological Specimen Preparation* 245–256 (SEM Inc., AMF O'Hara, Chicago, 1986).

10. Inoue, T. & Osatake, H. *J. Electron Microsc.* 33, 356–362 (1984); Robards, A. W. & Sleytr, U. B. in *Low-Temperature Methods in Biological Electron Microscopy* (ed. Glauret, A. M.) 5–130 (Elsevier, New York, 1985).
11. Kenjo, T., Arai, H. & Suzuki, T. *JEOL news* 25E, 13–17 (1987).
12. Smart, P. & Tovey, N. K. in *Electron Microscopy of Soils and Sediments* 26–111 (Clarendon, Oxford, 1981).
13. Smart, P. & Tovey, N. K. in *Electron Microscopy of Soils and Sediments* 50–55 (Clarendon, Oxford, 1982).
14. Adrian, M., Dubochet, J., Lepault, J. & McDowell, A. W. *Nature* 308, 32–36 (1984).
15. Iwatsuki, M. & Harada, Y. *J. Microsc. (Japan)* 21, 174–179 (1987).
16. Iwatsuki, M., Kihara, H., Nakanishi, K. & Harada, Y. *Proc. XIth Int. Cong. on Electron Microscopy Kyoto* 251–254 (1986).
17. Nakazawa, H., Yamada, H., Fujita, T. & Ito, Y. *Clay Sci.* 6, 269–276 (1987).

Tectonic controls on the uplift of the Nanga Parbat Massif, Pakistan Himalayas

Robert W. H. Butler* & David J. Prior†

*Department of Earth Sciences, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK

†Department of Earth Sciences, The University, Leeds LS2 9JT, UK

Cenozoic mountain belts like the Himalayas contain regions that experienced rapid uplift in comparison with the areas around them¹. Whether this uplift occurred by buoyant rebound of previously thickened crust² or during the crustal thickening process itself is uncertain^{3,4}. Here we show, however, that the 7 mm yr⁻¹ Recent uplift⁵ of the Nanga Parbat massif, Pakistan Himalayas, was accomplished during NW-directed thrusting located along its western margin. This 3-km-wide Liachar thrust zone operated from amphibolite facies to the present topographic surface and accommodates substantial crustal shortening. The adjacent Indus river valley, itself cut by seismogenic faulting, completely transects the massif, showing that erosion keeps pace with geologically rapid thrust uplift⁶.

The Nanga Parbat massif (Fig. 1) represents the exposed northern limit of Indian continental crust within the Himalayan collision belt^{7,8}. It lies within a structural half-window^{9,10} (the Nanga Parbat syntaxis) through the principal Tertiary suture, the Main Mantle thrust¹¹ (MMT), and is surrounded by the Kohistan arc terrane of the northern (Asian) block^{8,10}. During the collision process, probably in late Eocene times¹², the Nanga Parbat basement gneisses, migmatites and metasediments together with cover sediments were subducted beneath the Kohistan arc so that all units experienced *P-T* conditions suitable for amphibolite facies metamorphism^{10,13}. Deformation then migrated north and south away from the MMT so that the 1,500–2,000 km of continued plate convergence was accommodated in Tibet and the rest of the Himalayas¹⁴. Pioneering studies by Zeitler⁵ of fission track and ⁴⁰Ar/³⁹Ar mineral cooling ages from Kohistan and the northern Himalayas of Pakistan show that most of the suture zone experienced a gradual uplift of <6 km over the past 35 Myr. However, the Nanga Parbat massif has experienced >10 km uplift in the past 10 Myr, of which >6 km has occurred in the past 1.3 Myr. As such it is one of a series of recently uplifted parts of northern Pakistan, although it is the only one to bring lower (Indian) plate rocks to outcrop. The rapid uplift is indicated by the elevation of Nanga Parbat (8,126 m) at about twice the height of the ground 50 km west along the MMT.

A key to unravelling complex geological structures is to define the relative movement sense and its direction on faults and shear zones¹⁵. It is understood that small-scale geological structures record small increments of the displacement paths within fault and shear zones¹⁶. These can be documented from simple field observations on lithologies containing suitable objects such as rigid bodies¹⁷ or containing localized shears with recognizable bend-in fabrics^{18,19}. These types of structures can occur over a broad range of deformation mechanisms so that it is possible

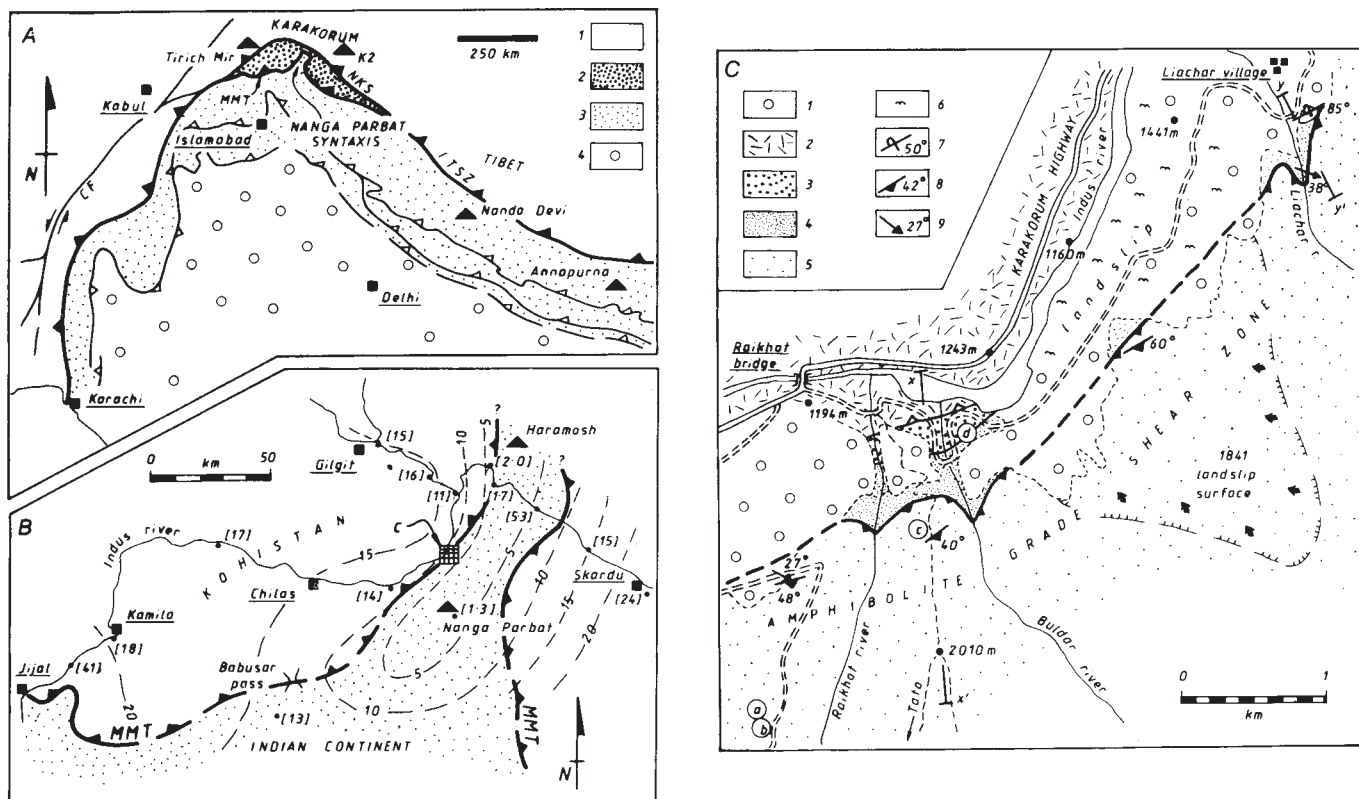


Fig. 1 A, Tectonic setting of the Nanga Parbat massif in the NW Himalayas^{7,21}. CF, Chaman fault; MMT, Main Mantle thrust; ITSZ, Indus Tsangpo suture zone; NKS, northern Kohistan suture; 1, Asian and Afghan terranes; 2, Kohistan arc terrane; 3, Himalayas and the stacked Indian continental thrust systems; 4, foreland. B, The Nanga Parbat massif. Stippled rocks are Indian continent, the Kohistan terrane is unornamented. Contours are fission track ages on zircon⁵ (measure at ~200 °C). C, Detailed map of the Raikhot-Liachar area on the western margin of the Nanga Parbat massif. 1, Indus valley fill (fluvio-glacial and debris flow deposits); 2, Kohistan arc material; 3, Nanga Parbat cover rocks; 4, Nanga Parbat gneisses with flattening fabrics; 5, Nanga Parbat gneisses with shear zone fabrics; 6, landslide debris from the 1841 event; 7, bedding; 8, mylonitic foliation; 9, linear fabrics in mylonites (movement direction); x-x', section line of Fig. 3a, y-y', section line of Fig. 3b. Locations of Fig. 2 shown as a, b, c, d.

to build up fault-plane solutions for ancient structures regardless of their seismic character.

To shed light on the uplift processes acting on the Nanga Parbat massif we present data collected during mapping and geological traverses executed near the Karakoram highway at Raikhot (Fig. 1C). The MMT along this part of the western margin of the massif is a distinct mappable zone of amphibolite-grade shear involving rocks of the Kohistan arc, an old metamorphic basement (the Nanga Parbat gneisses) and a thin veneer of younger metasedimentary cover²⁰. Regional criteria dictate that this shear zone was subhorizontal during displacement^{20,21} with small-scale linear structures and shear sense criteria indicating SSE-NNW subduction of the Indian continent under Kohistan. Subsequent thrust stacking and the uplift of the Nanga Parbat massif have rotated the MMT to vertical¹⁰.

About 10 km south of this early MMT, around the village of Tato (Fig. 1C), outcrops of Nanga Parbat basement tonalitic gneisses with pelitic and calc-metasediments, metabasic intrusives and granite veins show multiple tectonic and metamorphic fabrics with kyanite-bearing assemblages and migmatitic textures in pelitic units. Within ~2 km of the MMT between Tato and Raikhot the Nanga Parbat basement is predominantly tonalitic although all lithologies are penetratively deformed within a well-ordered moderately SE-dipping shear zone (Fig. 2a). Shear sense criteria indicate NW-directed overthrust movement of the Nanga Parbat massif back over Kohistan.

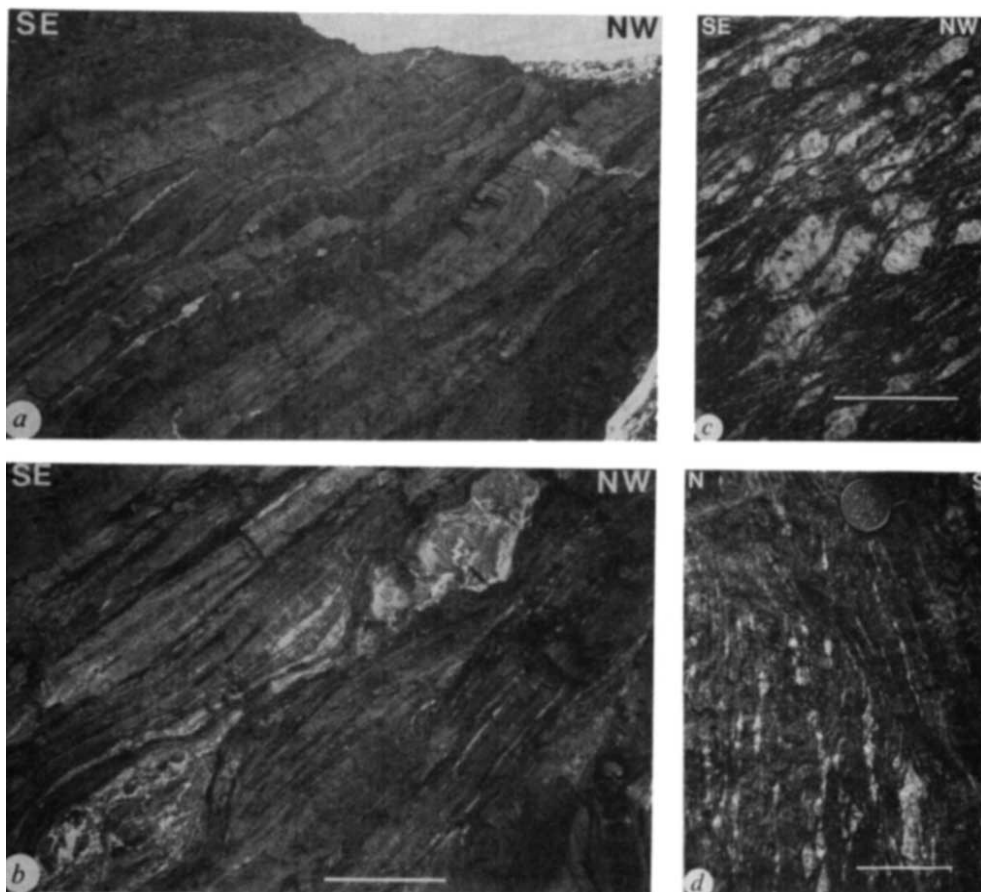
Granitic sheets within the gneisses show systematic reverse-sense offsets (Fig. 2a) on all scales. Feldspar megacrysts (~3 cm across) show pronounced recrystallized tails (Fig. 2c) with a dominant asymmetry; localized zones of deformation step down across slightly earlier foliation and layering (Fig. 2b, c) to define

shear bands²². NW movement axes are indicated by quartz-rod stretching lineations. The mesoscopically ductile deformation (for example Fig. 2a, b) appears to be controlled by crystal-plastic microscopic deformation mechanisms in quartz and feldspar. These observations, together with the widespread preservation of 'syn-kinematic' amphibolite facies metamorphic assemblages in the thrust-shear zone, suggest initiation and displacement at temperatures >450 °C. Penetrative deformation also occurs into upper greenschist (biotite) facies, but a more radical temperature drop during faulting is represented by rare 25 cm horizons of cataclases. These contain local riedel shears²³ indicative of NW-directed thrusting.

About 500 m above the Indus, south of Raikhot, the mesoscopically ductile, amphibolite-grade thrust-shear zone is carried on a 20-m-wide fault zone marked by micro-brecciation and gouges. This fault zone, named here the Liachar thrust, represents displacements at the highest crustal levels. It has an abrupt upper contact and caps a 500-m-wide zone of spaced NW- and SE-directed compressional cataclastic faults containing gouges indicative of low-temperature deformation. There is a tendency for increased SE-directed faulting away from the Nanga Parbat massif and into the Kohistan terrane. Critically all these gouge zones cut the MMT once it had attained a steep dip and no SE-directed faults were found above the Liachar thrust in the Nanga Parbat massif.

Among the high-level fault zones beneath the Liachar thrust there are indications of earlier deformation on ultra-fine-grained fault zones within Nanga Parbat tonalites. Foliation is vertical and bends into the shallower SE-dipping shears to indicate NW-directed thrusting (Fig. 2d). Feldspar megacrysts are not retrogressed in the shear zones (Fig. 2d), indicating deformation

Fig. 2 Shear criteria on the amphibolite-facies structures, Raikhot and Buldar areas. *a*, Deflection of leuco-granitic sheets intruded within metasediments of the Nanga Parbat gneisses; the cliff is ≈ 400 m high (*a* on Fig. 1C). *b*, localized shear band structure causing asymmetric boudinage of amphibolite sheet, Nanga Parbat gneisses. Scale bar 1 m (*b* on Fig. 1C). *c*, Asymmetric feldspar augen and small-scale shear bands in deformed tonalities, Nanga Parbat gneiss. Scale bar, 5 cm (*c* on Fig. 1C). *d*, Flattened tonalitic Nanga Parbat gneisses with local shear zone tip, footwall to Liachar cataclastic thrust zone. Scale bar, 5 cm (*d* on Fig. 1C).



at a much higher temperature than the gouge zones. However, very fine-grained zones probably composed of ultracataclasites are found in association with high-temperature shears and are thought to represent radical increases in strain rate rather than local temperature drops²⁴. Significant flattening is also evident in the tip regions of NW and SE-directed shears, although the whole region beneath the Liachar thrust shows megacryst shapes flattened in the foliation. This deformation apparently represents horizontal shortening before uplift and the generation of the high-level gouge zones with the Liachar thrust.

The 20-m gouge zone above Raikhot can be mapped as a continuous wide carpet of cataclasites derived from the Nanga Parbat gneisses (Figs 1C, 3b). In its type area of the lower Liachar valley this thrust zone overrides Quaternary Indus valley sediments which locally have acquired a vertical dip. Further evidence for very young movement on the Liachar thrust is indicated by the adjacent 5-km-long mountainside failure scarp, which was triggered by an earthquake in 1840.

In its type area the Liachar thrust zone consists of three major gouge zones linked by sub-faults showing either compressional or extensional geometries with displacements directed towards the NW quadrant (Fig. 3b). This pattern of faulting is strongly reminiscent of the riedel shear arrays²³ generated in small-scale gouge experiments. The lower major gouge zone shows ≥ 200 m offset over the Quaternary sediments but the sub-faults generally show < 1 m offsets. The sub-faults are marked by substantial wall-rock damage by penetrative fracturing. In contrast the three major gouge zones, which contain the vast bulk of the displacement on the Liachar thrust zone, show little wall-rock damage. These zones appear to have initiated by fracture processes leading to disaggregation, which permitted a change in deformation mechanism to grain-boundary sliding²⁵. So higher parts of the earthquake magnitude spectrum will have been dominated by the sub-faults showing their steep thrust and shallow normal fault motions whereas the lower, more aseismic parts will contain

the major low-angle thrust faults. Kinematic analyses defined by fault plane solutions from the few larger magnitude earthquakes in the absence of complete microseismic monitoring and field structural observations could be greatly misleading.

All the small-scale structures within the cataclastic Liachar thrust zone are indicative of NW-directed thrusting of the Nanga Parbat massif out onto Kohistan. As such the high-level displacements appear to form just the latest part of a thrusting history which initiated at ~ 12 km, under amphibolite facies conditions (Fig. 4). It appears from the Raikhot-Buldar area that part of this displacement was partitioned into net vertical pure shear, both under amphibolite grade and subsequently at higher crustal levels. These zones of pure shear are presumed to be responsible for much of the reorientation of the early MMT suture to its currently steep attitude. The SSE-directed shear criteria²⁰ associated with MMT movement are restricted to a narrow zone near this thrust and are generally modified by the pure shear strains. Otherwise all shear criteria, from amphibolite facies to high-level gouge fabrics in the Liachar thrust zone, indicate NW-directed uplift of the Nanga Parbat massif.

To explain the very young cooling ages of rocks now at the Earth's surface⁵, material must be removed from above the Nanga Parbat massif during movement on the Liachar thrust zone. Regional drainage patterns⁶ completely cross the massif (Fig. 1B) and so pre-date uplift. They currently transport sediment derived from the uplifting area out beyond the western margin of the massif and are accompanied by local catastrophic landslips. Thus the removal of ~ 6 km of rock from above Nanga Parbat in the past 1.3 Myr can be accomplished by erosional processes. The horizontal component of movement on the Liachar thrust probably lies between 15 and 30 km. Thus thrust probably initiated at amphibolite facies before the oldest $^{40}\text{Ar}/^{39}\text{Ar}$ age obtained from biotites within the massif⁵, namely 7 Myr. But the cooling history of the massif results both from uplift and heat loss by convective processes. Some of the cata-

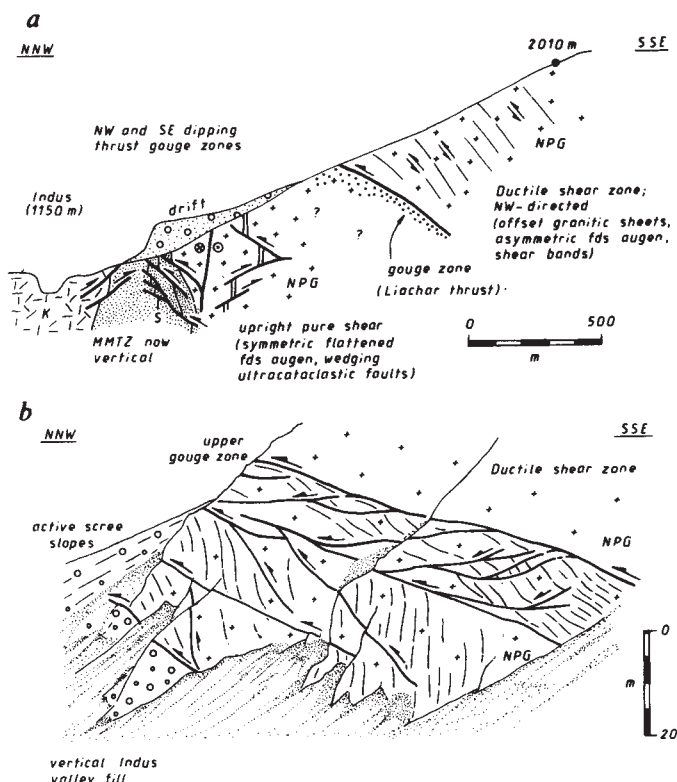


Fig. 3 a, Sketch section through the Buldar-Raikhot area (x-x' on Fig. 1C). Panel b, Field sketch of the Liachar thrust in its type area (y-y' on Fig. 1C).

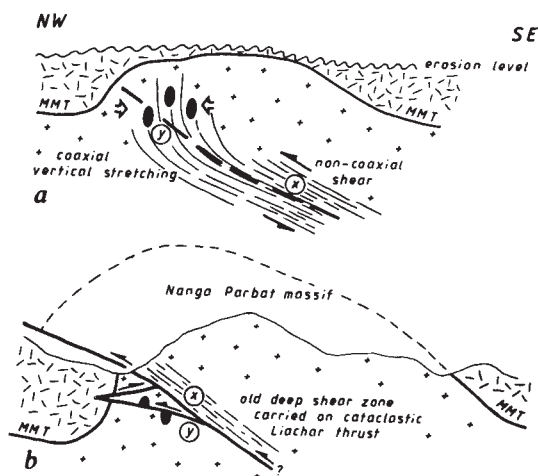


Fig. 4 Sequential evolution (a-b in time) of the Nanga Parbat uplift. a, The development of a simple shear zone (x) and a pure shear flattening (y) at a tip. The flattening zone rotates the MMT to near vertical. Movement through the system juxtaposes the two strain domains (b) eventually to break out across the Earth's surface in the Indus valley. Erosion of the Nanga Parbat massif occurred at a rate almost equal to that of the uplift. Sketches are true scale; the massif is ~30 km across.

clastic faults contain hydrothermally altered gouges, and hot springs mark the present outcrop of the Liachar thrust. Multidisciplinary studies involving linked metamorphic, isotopic, geochronological and structural analyses should elucidate the interplay between uplift and fluid processes so that at least the latter parts of the thermal evolution of the Nanga Parbat massif may be related to tectonic activity.

Fieldwork was funded by NERC. R.W.H.B. acknowledges a Royal Society (Jaffé Donation) Research Fellowship held in the

University of Durham. We thank colleagues there and at Leeds together with Lewis Owen (Leicester University) for discussions on Himalayan neotectonics.

Received 3 December 1987; accepted 7 March 1988.

- Gansser, A. in *Mountain Building Processes* (ed. Hsu, K. J.) 221-228 (Academic, New York, 1982).
- Richardson, S. W. & England, P. C. *Earth planet. Sci. Lett.* **42**, 183-190 (1979).
- Lyon-Caen, H. & Molnar, P. *J. geophys. Res.* **88**, 8171-8191 (1983).
- Ni, J. & Barazangi, M. *J. geophys. Res.* **88**, 1147-1163 (1984).
- Zeitler, P. K. *Tectonics* **4**, 127-151 (1985).
- Seeber, L. & Gornitz, V. *Tectonophysics* **92**, 335-367 (1983).
- Gansser, A. *Geology of the Himalayas* (Wiley, London, 1964).
- Tahirkehi, R. A. K. & Jan, M. Q. *Geol. Bull. Univ. Peshawar spec. Iss.* **11**, 1-189 (1979).
- Sarwar, G. & De Jong, K. A. in *Geodynamics of Pakistan* (eds Farah, A. & De Jong, K. A.) 341-349 (Geological Survey of Pakistan, Quetta, 1979).
- Coward, M. P. et al. in *Collision Tectonics* (eds Coward, M. P. & Ries, A. C.) (*Spec. Publ. geol. Soc. Lond.* 19), 203-219 (1986).
- Tahirkehi, R. A. K. et al. in *Geodynamics of Pakistan* (eds Farah, A. & De Jong, K. A.) 125-130 (Geological Survey of Pakistan, Quetta, 1979).
- Searle, M. P. et al. *Bull. geol. Soc. Am.* **98**, 678-701 (1987).
- Misch, P. *Contr. Miner. Petrol.* **10**, 315-356 (1964).
- Patriat, P. & Achache, J. *Nature* **311**, 615-621 (1984).
- White, S. H., Bretan, P. G. & Rutter, E. H. *Phil. Trans. R. Soc. A* **317**, 81-97 (1986).
- Cobbold, P. R., Gapais, D., Means, W. D. & Treagus, S. H. (eds) *J. struct. Geol.* **9**, 521-778 (1987).
- Platt, J. P. & Vissers, R. L. M. *J. struct. Geol.* **2**, 397-410 (1980).
- Brunel, M. *Tectonics* **5**, 247-265 (1986).
- Ramsay, J. G. & Graham, R. H. *Can. J. Earth Sci.* **7**, 786-813 (1970).
- Butler, R. W. H. & Prior, D. J. *Geol. Rdsch.* **77**, 239-255 (1988).
- Coward, M. P., Butler, R. W. H., Knipe, R. J. & Khan, M. A. *J. Geol. Soc. Lond.* **144**, 377-391 (1987).
- Simpson, C. & Schmid, S. *Bull. geol. Soc. Am.* **94**, 1281-1288 (1983).
- Chester, F. M., Friedman, M. & Logan, J. M. *Tectonophysics* **111**, 139-146 (1985).
- Sibson, R. H. *J. geol. Soc. Lond.* **133**, 191-213 (1977).
- Wojtal, S. & Mitra, G. *Bull. geol. Soc. Am.* **97**, 674-687 (1986).

Deposition of ungraded muds from high-density non-turbulent turbidity currents

I. N. McCave & K. P. N. Jones

Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

Turbidity current deposits of mud have recently been found up to 10 m in thickness resulting from single depositional events^{1,2}. Turbidites 1-5 m thick from the Madeira Abyssal Plain^{3,4} display long segments with little vertical change of grain size^{5,6}. This suggests that the currents 'froze' in some way, making them unable to sort sediment. Here we suggest that a decline in flow speed of these currents led to a concentration gradient in silt close to the bed that was sufficiently steep to stratify the flow. A concentration gradient of 18 kg m^{-3} over 1.5 m height gives a layer Richardson number above the critical value of 0.25 for damping of turbulence. We propose that, with turbulence severely damped, a dense suspension of fine material consolidates like a static suspension while continuing to flow downslope. The current changes from a fully turbulent flow to a decelerating flow with suppressed turbulence and increasing density. In the latter state interparticle forces prevent differential settling of the coarser grains.

We have analysed in detail a thick mud turbidite from 23 locations on Madeira Abyssal Plain with an X-ray scanning settling tube with digital recording and processing⁶ (Fig. 1). At several locations a well-sorted silt base several cm to tens of cm thick is overlain by >1 m of uniform mud (Fig. 1). The silt base shows upward fining in the modal size for diameters $<7\phi$ (that is, $>8 \mu\text{m}$). The transition from graded silt to uniform mud is fairly abrupt, occurring over 5 cm. The finer mud layer shows complex polymodal grain-size distributions. Two of the silt modes are prominent enough to correlate throughout the layer and are constant in size within the precision of the analysis ($\pm 0.1\phi$)⁶. The nearly uniform size of the mud layer is shown by several parameters: modal sizes, median and percentage silt.

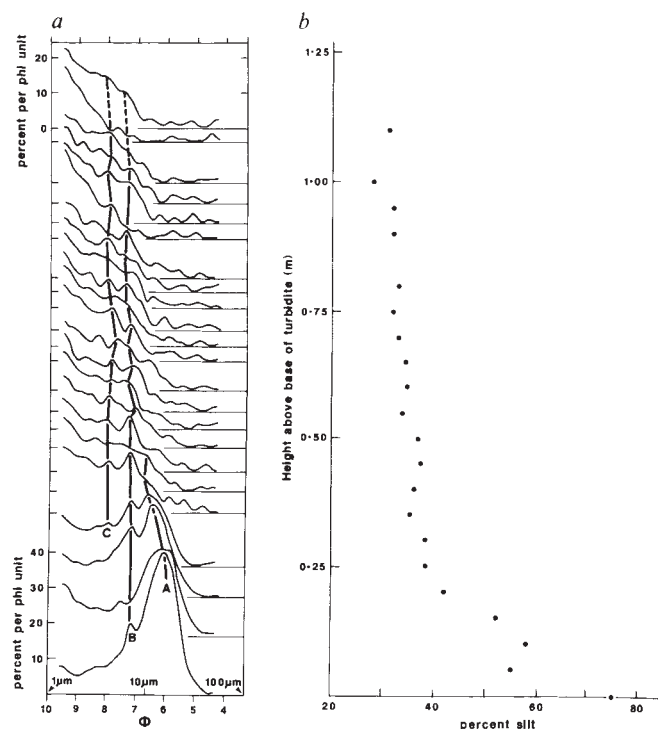


Fig. 1 Example of detailed grain-size analysis of uniform mud turbidite from the Madeira Abyssal Plain (Core D10971, sects 3 and 4): *a*, stacked frequency curves showing modes, A confined to the silt-rich base and fines up, and B and C which maintain constant size up to the top of the unit; *b*, percentage of silt ($2-63\ \mu\text{m}$), showing abrupt increase near bottom of unit.

The gradients of the Madeira Abyssal Plain are very low, the depositional relief being $\leq 30\text{ m}$ over hundreds of km (Fig. 2). The thickness and grain size of the silty base of one turbidite ('D' of Weaver *et al.*³) suggest inflow from the northeast, proceeding around several low abyssal hills. A variant lacking a silt base occurs most commonly at the distal southwest end of the flow path or in culs-de-sac or depressions where final slowing down or ponding of the flow are likely. The individual turbidites show an overall increase in thickness downcurrent, but a decline in thickness of the basal silt. Cores taken from up the side of an abyssal hill show turbidite D to thin by over 95% 15 m above the present level of the adjacent abyssal plain, suggesting a flow $\leq 20\text{ m}$ thick.

Stow and Bowen⁷ suggested that thin mud turbidites were deposited from low-concentration flows. In these, deposition by trapping in the viscous sublayer⁸ yields size segregation resulting in lamination⁷, a feature conspicuously absent from uniform mud turbidites (or 'homogenites' or 'unifites')^{2,9}. The deposition rate of aggregated particles of settling velocity 1 mm s^{-1} from a flow of low concentration (2 kg m^{-3}) is $<7\text{ kg m}^{-2}\text{ h}^{-1}$ ($<0.005\text{ m h}^{-1}$). A 5-m turbidite would require a flow lasting about 6 weeks. Such a flow duration is at the limit of credibility for a low-concentration flow across Madeira Abyssal Plain and beyond that limit for the 10-m-thick deposits in the Hellenic Trench^{1,2,5}.

We propose that these turbidity currents were of high concentration ($\sim 50-100\text{ kg m}^{-3}$) just before deposition, rather than low concentration ($\sim 2\text{ kg m}^{-3}$). This proposal stems from two recent observations: first, sorting is suppressed in mud suspensions of high concentration which develop cohesive strength¹⁰, and second, such concentrations are found in energetic shallow tidal waters, from which high-concentration suspensions known as fluid mud become segregated during flow¹¹⁻¹⁴. These two observations suggested to us that the analogy of fluid mud formation might be applied to a highly turbid underflow down a declining slope.

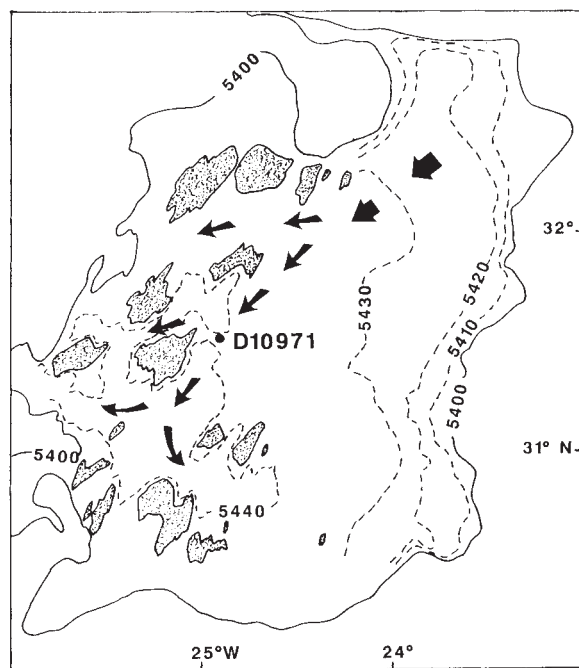


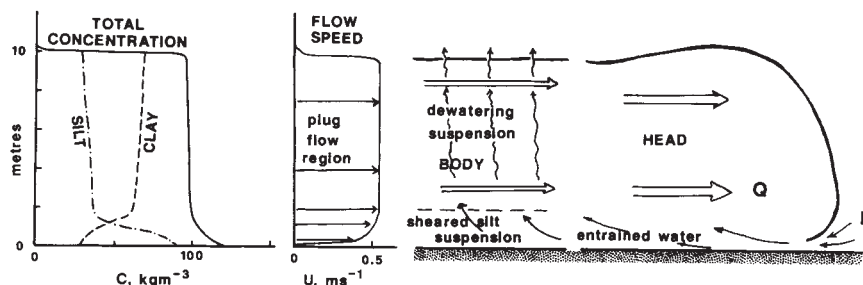
Fig. 2 Bathymetry (m) of part of Madeira Abyssal Plain in which thick mud turbidites are found. Path of turbidity current D shown by arrows. Location of core containing turbidite analysed in Fig. 1 also shown. Stippled areas denote abyssal hills which rise $\sim 200\text{ m}$ above the plain.

Size sorting was suppressed at concentrations $>140\text{ kg m}^{-3}$ in Been and Sills' settling experiments with clayey silt (30% clay) in tap water¹⁰, and at 50 kg m^{-3} in size measurements by Stein on montmorillonitic clays with dispersant¹⁵. For fully flocculated suspensions the flocs can fill the space available at solid-particle volume concentrations as low as 0.01 (25 kg m^{-3}) for platy clay minerals^{16,17}. Above this concentration, sedimentation changes from discrete settling to network collapse and suspensions display non-linear elastico-viscous behaviour. The network collapse regime shows up as a sharp density step in settling suspensions, a feature observed both in laboratory settling columns¹⁰ and in the Severn Estuary, where dense suspensions formed while such a density step lowered at 1 mm s^{-1} for several hours in a tidal current slowing from 0.60 to 0.15 m s^{-1} (ref. 13).

As in most cases with deep-sea turbidity currents, properties of the flow must be inferred from the deposit. The further one goes back upstream from the site of deposition, the greater the speculation involved. We suggest that such a flow started as a slump. It must then have evolved into a highly turbulent current to provide stresses high enough to disaggregate the sediments completely. As the seabed gradient decreased along the flow path, the current slowed down and evolved into a higher-density flow by particle settling as in the estuarine case mentioned above. Just before it stopped, the flow comprised a lower silty shear zone in which the concentration gradient was such as to damp turbulence, leaving low-frequency mixing capable of keeping 6ϕ silt (settling velocity 0.015 cm s^{-1}) in suspension. Above that the bulk of the flow was a clay-rich fluid mud with only sluggish overturn because of its cohesive strength and high viscosity. This behaviour has been reported from flume experiments on hyperconcentrated flow¹⁸. In these experiments with high concentrations the flow developed two zones, an upper one of uniform velocity and negligible turbulence and a lower-velocity-gradient zone in which turbulent intensity and median fluctuation frequency decreased with increasing sediment concentration¹⁸. Physical aspects of the proposed flows are outlined below.

In its early stages the flow down the Madeiran continental

Fig. 3 Sketch of fluid-mud turbidity current, showing head region with flux Q and entrained flux F of clear water, body region with basal sheared silt suspension overlain by settling clay-rich plug of mud, velocity profile (U), and concentration profiles (C) of the total suspension and its silt/clay components.



margin would have accelerated and become more dilute by entrainment. In the later stages, which concern us here, it became slower and increased in concentration. From a simple Chezy-type equation with the drag coefficient for a fine-grained slurry in pipe flow^{19,20} ($C_D \approx 4 \times 10^{-3}$), a flow of concentration 100 kg m^{-3} (density $\rho_t = 1,108 \text{ kg m}^{-3}$) 10 m thick over the abyssal plain of slope 1:5,000 has a steady speed of 0.65 m s^{-1} . A similar speed is obtained for a 20-m-thick flow of concentration 50 kg m^{-3} . The lower well-sorted silt-rich part of the flow arises from break-up of flocs by shear^{21,22} and flushing of the clay into the upper clay-rich zone (Fig. 3). Turbidity currents entrain some fluid at the head by overriding clear fluid ahead of the flow²³. This provides a continuous flux of fluid to flush clay out of the basal zone. This flux (F) is $< 0.13Q$, where Q is the volume flux of the whole flow downslope²³. We expect the principal accumulation of silt in the basal zone to have occurred while the flow was highly turbulent before the final fluid-mud development.

We argue that the concentration gradient of suspended silt near the bed was sufficient to stratify the flow and give the basal silty layer a Richardson number above the critical value of 0.25 required to damp turbulence²⁴. The concentration gradient under hyperconcentrated conditions is obtained by engineering approximations²⁵. With u_* , from the flow speed and drag coefficient, equal to 0.04 m s^{-1} , the exponent in the suspended-load equation modified for damped turbulence²⁵ is ~ 0.0195 , which gives an increase in concentration from 100 kg m^{-3} at 1.5 m above the bed to 118.4 kg m^{-3} at 1 mm. This concentration gradient yields a bulk Richardson number ($g\Delta\rho h/\rho_t\Delta U^2$) for the basal silty layer of 0.33, above the critical value. The concentration in the silty zone is higher than would be achieved under a steady turbulent shear flow over a silt bed. Using Smith's²⁶ expression for the concentration C_b in an equilibrium bedload layer we obtain $C_b = 3.6 \text{ kg m}^{-3}$. The conditions we propose arise because the flow is decelerating, and silt grains coming from the break-up of aggregates accumulate in the basal layer. It is not deposited immediately because the flow speed is above the $\sim 0.4 \text{ m s}^{-1}$ needed for deposition from a silt slurry²⁷. Thus we contend that the basal silt-rich layer stratifies the flow and suppresses turbulence leading to absence of active mixing at the base of the flow.

Finally, to develop higher concentration the flow must not entrain fluid at its upper boundary. For this to be the case the local or gradient Richardson number across the interface should also be more than about 0.25. Judging from modern fluid mud layers, the upper boundary of the turbidity current would be fairly sharp; we assume a thickness of 0.2 m here. For a flow of concentration 100 kg m^{-3} and speed 0.65 m s^{-1} , $Ri = 0.24$. Thus the parameters are of the correct order to prevent dilution of the flow by entrainment at the upper boundary.

What we have outlined here is a process by which a decelerating turbulent flow evolves into a type of hyperconcentrated flow or debris flow²⁸. It may also be applicable to the formation of fluid mud in estuaries. Unlaminated mud deposits of uniform size are commonly much thinner than the turbidites described here²⁹. Deposition from thinner hyperconcentrated layers may also be a relatively common mode of mud deposition.

We are grateful to Drs P. P. E. Weaver of the Institute of

Oceanographic Sciences and R. T. E. Schuttenhelm of the Rijks Geologische Dienst for access to the Madeira Abyssal Plain cores.

Received 17 December 1987; accepted 9 March 1988.

- Blanpied, C. & Stanley, D. J. *Smithsonian Contrib. mar. Sci.* **13**, 1-40 (1981).
- Cita, M. B. *et al. Mar. Geol.* **55**, 79-102 (1984).
- Weaver, P. P. E., Searle, R. C. & Kuijpers, A. *Spec. Publ. geol. Soc. Lond.* **21**, 131-143 (1986).
- Weaver, P. P. E. & Rothwell, R. G. *Spec. Publ. geol. Soc. Lond.* **31**, 71-86 (1987).
- Hieke, W. *Mar. Geol.* **55**, 63-78 (1984).
- Jones, K. P. N., McCave, I. N. & Patel, P. D. *Sedimentology* **35**, 163-172 (1988).
- Stow, D. A. V. & Bowen, A. J. *Sedimentology* **27**, 31-46 (1980).
- McCave, I. N. & Swift, S. A. *Bull. geol. Soc. Am.* **87**, 541-546 (1976).
- Stanley, D. J. *Geo-mar. Lett.* **1**, 77-83 (1981).
- Been, K. & Sills, G. C. *Géotechnique* **71**, 519-535 (1981).
- Allen, G. P. thesis, Univ. Bordeaux (1972).
- Kirby, R. & Parker, W. R. *Can. J. Fish. Aquat. Sci.* **40** (suppl. 1), 83-95 (1983).
- Parker, W. R. *Continental Shelf Res.* **7**, 1285-1293 (1986).
- Wells, J. T. *Can. J. Fish. Aquat. Sci.* **40** (suppl. 1), 130-142 (1983).
- Stein, R. J. *sedim. Petrol.* **55**, 590-593 (1985).
- Hunter, R. J. *Adv. Coll. Interface Sci.* **17**, 197-211 (1982).
- Williams, D. J. A. in *Estuarine Cohesive Sediment Dynamics* (ed. Mehta, A. J.) 110-125 (Springer, Heidelberg, 1986).
- Komar, P. D. In *The Sea Vol. 6 (Marine Modelling)* (eds Goldberg, E. D. *et al.*) 603-621 (Wiley, New York, 1977).
- Woodcock, C. R. & Mason, J. S. *Bulk Solids Handling*, 128-150 (L. Hill, Glasgow & London, 1987).
- Wang, M., Zhan, Y., Liu, J., Duan, W. & Wu, W. In *Proc. 2nd Int. Symp. River Sedimentation* 45-53 (Water Resources and Electrical Power Press, Nanjing, 1983).
- Hunt, J. R. in *Estuarine Cohesive Sediment Dynamics* (ed. Mehta, A. J.) 85-109 (Springer, Heidelberg, 1986).
- McCave, I. N. *Spec. Publ. geol. Soc. Lond.* **14**, 35-69 (1984).
- Simpson, J. E. A. *Rev. Fluid. Mech.* **14**, 213-234 (1982).
- Turner, J. S. *Buoyancy Effects in Fluids* (Cambridge University Press, 1973).
- van Rijn, L. C. *J. Hydraul. Engng.* **110**, 1613-1641.
- Smith, J. D. In *The Sea Vol. 6 (Marine Modelling)* (eds Goldberg, E. D. *et al.*) 539-577 (Wiley, New York, 1977).
- Sinclair, C. G. in *Interaction Between Fluids and Particles* 78-86 (Inst. Chem. Engrs., London, 1962).
- Pierson, T. C. & Costa, J. E. *Rev. Engng. Geol.* **7**, 1-12 (1987).
- Stow, D. A. V. & Piper, J. D. W. *Fine Grained Sediments: Deep Water Processes and Facies. Spec. Publ. geol. Soc. Lond.* **15**, 659 pp.

Palaeoenvironmental trends in the history of trace fossils

David J. Bottjer*, Mary L. Droser† & David Jablonski‡

* Department of Geological Sciences, University of Southern California, Los Angeles, California 90089, USA

† Department of Geology, Oberlin College, Oberlin, Ohio 44074, USA

‡ Department of Geophysical Sciences, University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637, USA

Tracks, trails and burrows made by organisms are commonly preserved in Phanerozoic sedimentary rocks. These biogenic sedimentary structures, or trace fossils, have been most intensively studied from rocks deposited in marine environments, and have proven to be extremely useful in palaeoenvironmental reconstructions. Our systematic analysis of occurrences in the stratigraphic record of two of the most common trace fossils, *Zoophycos* and *Ophiomorpha*, indicates that occurrences for each show an environmentally linked directional pattern through time. *Zoophycos* is commonly reported from Palaeozoic strata deposited from near-shore to deep-sea environments, whereas in post-Palaeozoic strata

it is generally reported only from rocks deposited in deeper-water environments. *Ophiomorpha* first appeared in Permian shallow-water environments and was common in deep-sea fan environments by the Cretaceous, and has been distributed from nearshore to deep-sea environments since then. These onshore-offshore trends, which are similar to those previously reported for Phanerozoic benthic macroinvertebrate assemblages and individual higher taxa, require integration into future palaeoenvironmental reconstructions that use trace fossils.

To investigate onshore-offshore trends for *Zoophycos* and *Ophiomorpha*, we have surveyed the literature of global Phanerozoic trace fossil occurrences. We have used the commonly accepted definitions of these two trace fossil ichnogenera (binary nomenclature—ichnogenus and ichnospecies—is used to classify trace fossils)¹. Data were plotted in a modified version of the time-environment diagram introduced by Sepkoski and Sheehan (ref. 2, see also refs 3 and 4). As before, time in millions of years (Myr) is plotted on the vertical axis, and position within an environmental transect, from marine nearshore to slope and deep basin, is on the horizontal axis; ichnogenera are recorded within this time-environment matrix as presence-absence. Due to the imprecise nature of dating of much of the Phanerozoic stratigraphic record, we have divided the time axis into 10-Myr intervals. A given sedimentary unit was correlated to stratigraphic stage or substage on the basis of published reports, and assigned to a time increment using the time scale of Harland *et al.*²; use of other time scales does not alter results.

The horizontal axis of the diagram has five palaeoenvironmental partitions, based on the gradient of physical processes observed in most marine subtidal nearshore to deep-basin environments (see Table 1). Trace fossil occurrences were placed within this gradient primarily through analysis of associated bedding features and physical sedimentary structures (Table 1). Comparative taphonomic data (for example, ref. 6) was also used for environmental placement. These criteria and other less general indicators (see chapters in refs 7, 8), along with stratigraphic context within facies sequences, were used to determine the broad environmental setting of all the trace fossil data incorporated into each time-environment diagram. Note that our palaeoenvironmental framework is based on physical energy levels, and not geographic position of environments, so that, for example, a barrier bar is classified as nearshore, a lagoon is classified as inner shelf. But a geographic component does exist in this type of analysis in that our nearshore and inner shelf data points would always be located shorewards of our middle shelf, outer shelf, and slope and deep basin data points.

The palaeoenvironmental scheme provides a framework into which previously published information, heterogeneous as to level of detail and descriptive quality, can be fitted in a reasonably objective way. Our experience with the data currently available in the literature indicates that finer subdivision of this palaeoenvironmental scheme is at present impractical. We rely on the most commonly reported sedimentary-stratigraphic reasoning that would arise from incorporating palaeoecological analyses of fossil content. Such standardization of palaeoenvironmental subdivisions is clearly desirable for further studies of onshore-offshore histories.

Zoophycos, a complex spreiten structure with two basic forms (helicoidal and planar¹), occurs through most of the Phanerozoic (Fig. 1). Our oldest *Zoophycos* data point (480–490B) is from Lower Ordovician strata deposited in inner shelf environments. But because data are sparse for this time interval, this apparent environment of first occurrence should be viewed with caution. By the Early Silurian, *Zoophycos* was present in slope and deep basin environments (430–440E), and was present in nearshore environments by the Early Devonian (390–400A). *Zoophycos* was fairly common in nearshore habitats through the remainder of the Palaeozoic, after which time it is unknown from these environments (Fig. 1). The youngest inner and middle shelf occurrences (80–90B, C) are Late Cretaceous in age and the

youngest outer shelf occurrence (20–30D) dates from the Oligocene (Fig. 1). *Zoophycos* has remained common in slope and deep basin environments since its first Palaeozoic occurrence in that setting (Fig. 1).

Ophiomorpha, a three-dimensional branching burrow system with pelleted burrow linings¹, occurs primarily in the Mesozoic and Cenozoic (Fig. 2). The oldest known occurrence, however, is in Lower Permian strata deposited in nearshore environments (270–280A). By the Late Jurassic *Ophiomorpha* was present in inner shelf environments (150–160B), and was present in middle shelf to slope and deep basin environments by the mid-Cretaceous (90–100C, E) (Fig. 2). Since then it generally has occurred in all the studied environments, although it has remained most common in its original nearshore habitat (Fig. 2).

Perhaps the best known geologic application of trace fossils is as palaeoenvironmental indicators. Seilacher⁹ demonstrated that certain suites of trace fossils, characterized by similar trace morphology and hence tracemaker behaviour, typically recur in strata deposited under similar depositional conditions. Each characteristic suite is an *ichnofacies*, and each ichnofacies is named for a typical component trace fossil. These generally include the *Trypanites* (hard substrata), *Glossifungites* (firm substrata), *Skolithos* (nearshore shifting substrata), *Cruziana* (shelf above storm wave base), *Zoophycos* (outer continental shelf and slope), and *Nereites* (deep sea) marine ichnofacies^{10–12}. The ichnofacies concept is generally applied in palaeoenvironmental analysis by comparing the trace fossil content of a sedimentary unit to the suites of traces considered characteristic of each ichnofacies. The unit in question is then inferred to have been deposited under conditions similar to the conditions typical of the ichnofacies with the closest matching trace fossil composition.

The discovery of trace fossils in strata representing environments other than those of their typical ichnofacies has been considered troublesome, but has not been assessed in any systematic way. These exceptions to distributions predicted by ichnofacies models have been generally attributed to unusual occurrences of physical/biological conditions commonly associated with other environments^{10–12}. For example, shallow-water occurrences of *Zoophycos* could be attributed to the nearshore deposition of muds and muddy sands more typical of deep-water habitats. Recognition that such exceptions exist has led to the conclusion that local environmental conditions are most important in controlling the distribution of tracemakers (for example,

Table 1 Palaeoenvironmental scheme used in time-environment diagrams, with most important criteria for recognition

A, Nearshore: subtidal but above fairweather wave base. Criteria: (1) Thick laterally-continuous beds of sand and (commonly) coarser sediments with parallel laminae and cross-bedding. (2) Fossils disarticulated, abraded, well-bedded.
B, Inner shelf: below fairweather wave base, above normal storm wave base. Criteria: (1) Common storm beds with parallel laminae or hummocky-cross-stratification interbedded with fairweather fine-grained beds. (2) Fossils commonly in storm beds.
C, Middle shelf: below normal storm wave base, above maximum storm wave base. Criteria: (1) Massive fine-grained beds, commonly rhythmic, storm beds relatively uncommon. (2) Fossils unabraded.
D, Outer shelf: no storm influence, landward of shelf edge. Criteria: (1) Massive fine-grained strata, commonly rhythmic on fine scale, no evidence of slumping, mass movement, or turbidites. (2) Fossils unabraded.
E, Slope and deep basin: beyond shelf edge. Criteria: (1) Slumping, mass movements, turbidites, deep-sea fan facies geometries. (2) Fossils from beds deposited as turbidites or by some other type of mass movement omitted.

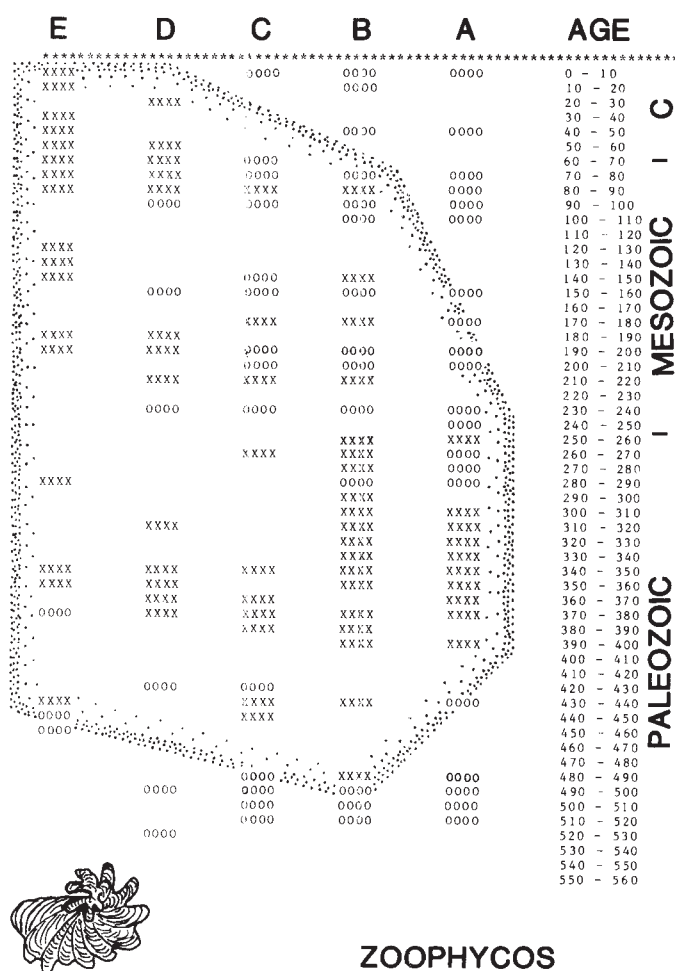


Fig. 1 *Zoophycos* time-environment diagram. Plan view sketch of *Zoophycos* specimen (greatest width is 20 cm) in lower left corner (modified from ref. 11). Vertical axis is in millions of years; C is Cenozoic. Horizontal axis shows palaeoenvironmental categories, from nearshore (A) to slope and deep basin (E) (definitions in Table 1). Presence of *Zoophycos* indicated by XXXX, absence of *Zoophycos* indicated by 0000. Area delineated by stippling shows time-environmental distribution of *Zoophycos*. Most of the literature sources for these data points can be found in Bottjer *et al.*²⁰; a complete list may be obtained from the authors upon request.

refs 10-12). But because any of these conditions commonly change progressively with water depth, recognition of ichnofacies is still considered useful in determinations of palaeobathymetry.

Our results on *Zoophycos* show at least a 150-Myr history of common occurrence in all environments that we examined until its disappearance from nearshore environments at the end of the Palaeozoic, followed by subsequent disappearance from shelf environments in the Cretaceous and Cenozoic. The Neogene and present occurrence of *Zoophycos* in slope and deep basin settings (Fig. 1) conforms to the environmental conditions defined for the *Zoophycos* ichnofacies. Interestingly, there have been previous indications in the literature that an onshore-offshore pattern exists for *Zoophycos*^{10,13}.

Conversely, over a period of 170 Myr *Ophiomorpha* progressively appeared in more offshore environments, from the oldest known occurrence in the Early Permian nearshore to its mid-Cretaceous appearance in slope and deep-basin settings. Therefore, *Ophiomorpha*, typically thought to be a common

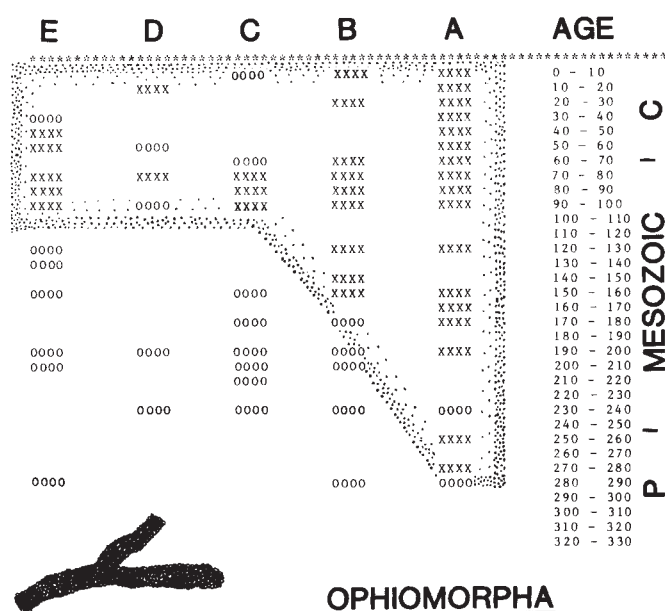


Fig. 2 *Ophiomorpha* time-environment diagram. Plan view sketch of *Ophiomorpha* specimen (lower branch length is 24 cm) in lower left corner (modified from ref. 19). Vertical axis is in millions of years; P is Paleozoic, C is Cenozoic. Horizontal axis shows palaeoenvironmental categories, from nearshore (A) to slope and deep basin (E) (for definitions, see Table 1). Presence of *Ophiomorpha* indicated by XXXX, absence of *Ophiomorpha* indicated by 0000. Area delineated by stippling shows time-environmental distribution of *Ophiomorpha*. Most of the literature sources for these data points can be found in Bottjer *et al.*²⁰; a complete list may be obtained from the authors upon request.

component of the *Skolithos* ichnofacies, was restricted to environments with bathymetric characteristics of this ichnofacies only during the first two-thirds (120 Myr) of its history (Fig. 2). All *Ophiomorpha* occurrences which we examined are found in sandstone, indicating that the organisms which made them widened the range of sandy substrata that they could colonize through time, from nearshore to shelf sands and then to submarine canyon and deep-sea fan sands.

Such onshore-offshore trends have been demonstrated for other trace fossils, although in a less systematic way. The oldest occurrences of *Palaeodictyon* and *Squamodictyon* are in Lower Cambrian strata deposited in shallow water environments, but later occurrences are restricted to deep-sea deposits¹⁴. *Scolicia*, traces attributed to burrowing heart urchins, first occur in nearshore and shelfal strata in the Jurassic and do not appear in deep-sea fan strata until the Cretaceous^{12,15}. Seilacher¹⁶ has added a community-level dimension to this pattern, by suggesting that tracemakers in fine-grained pre-turbidite sediments evolved *in situ*, whereas the post-turbidite community consists mainly of tracemakers that have immigrated from shelf environments. More generally, Crimes¹⁷ demonstrated that the earliest trace fossils occur in strata deposited in shallow water environments, and that trace fossils appear in strata deposited in deeper-water environments only later.

Through examination of benthic macroinvertebrate assemblages in different parts of the geologic column, a number of workers have found that relatively advanced or modern community types tend to appear first in nearshore settings and subsequently spread offshore, whereas offshore settings tend to harbour community types of more archaic aspect²⁻⁴. Studies of individual post-Palaeozoic higher taxa indicate that, in this interval at least, the onshore-offshore trend for community data

is actually underlain by individualistic clade histories that only appear to act together when viewed on a coarse time scale¹⁸. These analyses document two different environmental patterns for taxa that have originated onshore¹⁸. Migration offshore with persistence of representatives onshore is termed 'expansion'. Movement offshore with loss of onshore representatives is termed 'retreat'. This same terminology can be applied to the two trace fossil examples described herein. *Zoophycos* may show expansion in the Palaeozoic (as suggested, but not demonstrated, by the sparse early Palaeozoic data), and is followed by retreat in the Mesozoic and Cenozoic. *Ophiomorpha*, after late Palaeozoic nearshore origination shows expansion through the Mesozoic.

Onshore-offshore patterns can provide a better understanding of tracemaker characteristics for traces in which the tracemaker is presently unknown. For example, as clades tend not to move across the shelf at similar times and rates¹⁸, ichnogenera produced by a wide variety of higher taxa might be least likely to show an onshore-offshore pattern, whereas ichnogenera produced by members of one or two higher taxa might be more likely to have an onshore-offshore trend as part of their time-environment history. Thus *Ophiomorpha*, thought to be produced only by decapod crustaceans¹⁹, has a strong onshore-offshore history of expansion, as do typical heart-urchin burrows^{12,15}. Little is known about the nature or number of tracemakers that have produced *Zoophycos* (for example, see ref. 1). However, because it shows a significant onshore-offshore trend, it too may be produced by only one or a few higher taxa.

Although the observations underlying the ichnofacies concept generally remain undisputed, a number of ichnogenera show a more dynamic historical pattern. Integration of onshore-offshore trends of trace fossils with the ichnofacies concept thus requires the recognition that, although the broad behavioural trends reflected in morphological characteristics of trace fossil suites can be generally used for determinations of environment of deposition, particular ichnogenera could have a Phanerozoic history independent of that predicted by the ichnofacies concept.

We are grateful to the staff at all the libraries at which we have worked, in particular those at: University of Southern California, Hancock Library (especially Suzanne Henderson); University of California, Los Angeles, Geology and Geophysics Library (especially Nantawan S. McLeod); Field Museum of Natural History Library; and John Crerar Library, University of Chicago (especially Kathleen Zar). We acknowledge the Petroleum Research Fund, administered by the American Chemical Society, for support of this research.

Received 14 December 1987; accepted 16 March 1988.

1. Häntzschel, W. in *Treatise on Invertebrate Paleontology*, Pt. W, *Miscellaneous, Supp. 1* (ed. Teichert, C.) 269 (Geol. Soc. Amer. and Univ. Kansas, Lawrence, 1975).
2. Sepkoski, J. J. Jr & Sheehan, P. M. in *Biotic Interactions in Recent and Fossil Benthic Communities* (eds Tevesz, M. J. S. & McCall, P. L.) 673-718 (Plenum, New York, 1983).
3. Jablonski, D., Sepkoski, J. J. Jr, Bottjer, D. J. & Sheehan, P. M. *Science* **222**, 1123-1125 (1983).
4. Sepkoski, J. J. Jr & Miller, A. I. in *Phanerozoic Diversity Patterns* (ed. Valentine, J. W.) 153-190 (Princeton University Press, NJ, 1985).
5. Harland, W. B. et al. *A Geologic Time Scale* (Cambridge University Press, 1982).
6. Brett, C. E. & Baird, G. C. *Palaio* **1**, 312-327 (1986).
7. *Sedimentary Environments and Facies* (ed. Reading, H. G.) (Blackwell, Oxford, 1986).
8. *Facies Models* (ed. R. G. Walker) (Geoscience Canada Reprint Series, 1, 1984).
9. Seilacher, A. *Mar. Geol.* **5**, 413-428 (1967).
10. Frey, R. W. & Pemberton, S. G. *Bull. Can. petrol. Geol.* **33**, 71-115 (1985).
11. Ekdale, A. A., Bromley, R. G. & Pemberton, S. G. *Ichnology, The Use of Trace Fossils in Sedimentology and Stratigraphy* 317 (Soc. econ. Paleont. Miner. Short Course No. 15, Tulsa, Oklahoma, 1984).
12. Frey, R. W. & Seilacher, A. *Lethaia* **13**, 183-207 (1980).
13. Seilacher, A. in *Evolution of Animal Behavior* (eds Nitecki, M. H. & Kitchell, J. A.) 62-87 (Oxford University Press, 1986).
14. Crimes, T. P. & Anderson, M. M. *J. Paleont.* **59**, 310-343 (1985).
15. Smith, A. B. & Crimes, T. P. *Lethaia* **16**, 79-92 (1983).
16. Seilacher, A. *Acta geol. Hist.* **16**, 39-44 (1981).
17. Crimes, T. P. *Nature* **248**, 328-330 (1974).
18. Bottjer, D. J. & Jablonski, D. *Palaio* **3** (in the press).
19. Frey, R. W., Howard, J. D. & Pryor, W. A. *Palaeoogeogr. Palaeclimatol. Palaecol.* **23**, 199-222 (1978).
20. Bottjer, D. J., Droser, M. L. & Jablonski, D. in *New Concepts in the Use of Biogenic Sedimentary Structures for Palaeoenvironmental Interpretation* (ed. Bottjer, D. J.) 57-65 (Pacific Section Soc. econ. Paleont. Miner. Volume and Guidebook No. 52, Los Angeles, California, 1987).

Geoid roughness and long-wavelength segmentation of the South Atlantic spreading ridge

Dominique Gibert* & Vincent Courtillot†

* IFREMER Centre de Brest, BP 70, 29263 Plouzané, France

† Institut de Physique du Globe, 4 place Jussieu, 75232 Paris cedex 5, France

The configuration of accreting plate boundaries tends to reflect the geometry of the original continental break-up, yet there are also many indications that the detailed pattern of spreading-ridge and transform-fault segments may be an inherent part of the spreading process itself^{1,2}. The Mid-Atlantic Ridge, for example, is segmented, with a typical spacing of ~40 km between transform faults^{3,4}. Here, we use the short-wavelength (25-110 km) part of the Seasat altimetry data set to derive a map of the 'geoid roughness' in the South Atlantic between 1° N and 39° S. This map reflects changes in the local amplitude (envelope) of the geoid signal primarily associated with topography. Zones of high roughness correlate with major fracture zones, aseismic ridges and the Mid-Atlantic Ridge. The most remarkable feature is that the roughness is not uniform along the ridge but displays quasi-periodic variations with a wavelength of ~400 km, a segmentation an order of magnitude larger than described so far, which has persisted for at least 60 Myr. This segmentation may reflect either fabric generated at the ridge by normal plate tectonic processes^{1,5}, or small-scale convective instabilities rising from ~80-km-deep region of the upper mantle^{3,6,7}.

The South Atlantic (Fig. 1) is still one of the least known oceans, and the geoid data collected over it during the Seasat mission⁸ may constitute the most extensive and homogeneous

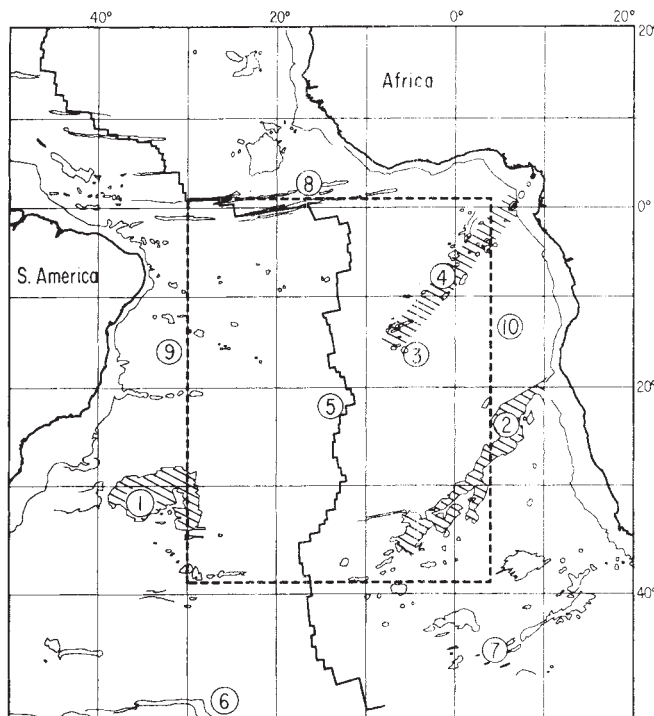


Fig. 1 Localization of the area of investigation (dashed line) in the South Atlantic. The geological units cited in the text are: 1, Rio Grande Rise (shaded); 2, Walvis ridge (shaded); 3, St Helena Island; 4, St Helena-Sao Tome-Principe line (shaded); 5, Mid-Atlantic Ridge; 6, Falkland fracture zone; 7, Agulhas fracture zone; 8, Equatorial fracture zones; 9, Brazil basin; 10, Angola basin, Mercator projection.

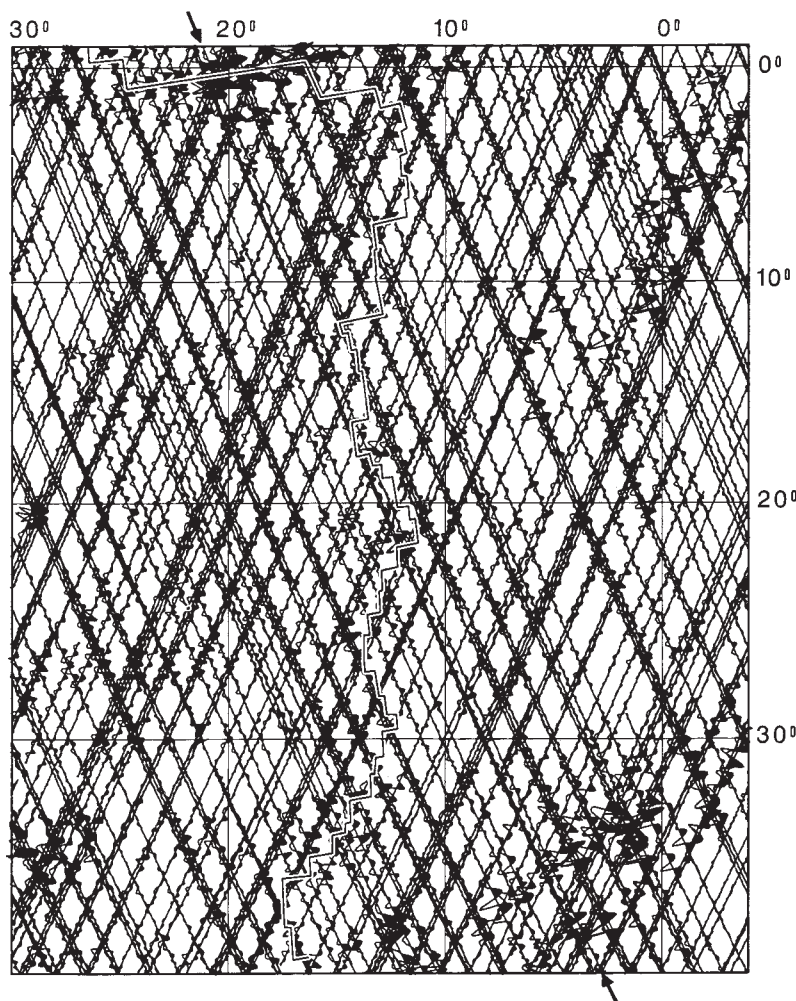


Fig. 2 High-pass-filtered (25–110 km) Seasat profiles available in the central part of the South Atlantic. Negative anomalies are black. The scale is such that a geoid anomaly of 1 m has an amplitude of 1° in longitude. The Mid-Atlantic Ridge is also shown for reference. Both ascending (northwestern run profiles) and descending (southwestern run profiles) tracks constitute a dense diamond-shaped net with a typical mesh size of 150 km. Along-track sampling interval is 7 km and the altimeter noise is ~ 7 cm r.m.s. Mercator projection.

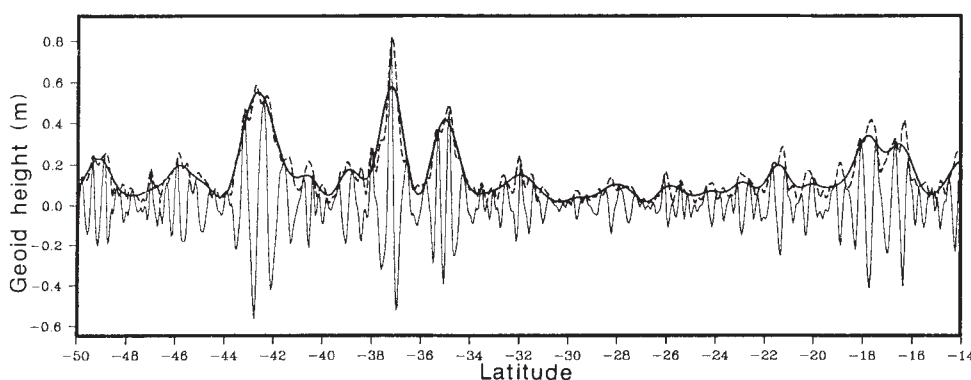


Fig. 3 Typical Seasat profile (shown by arrows in Fig. 2) processed according to the numerical procedure used to construct the geoid roughness map shown in Fig. 4. The envelope (dashed line) of the original high-pass-filtered (25–110 km) profile (thin solid line) is low-pass filtered (cutoff at 180 km, thick upper solid line). The resulting low-pass-filtered envelope displays the regional variations of the amplitude of the short-wavelength undulations of the geoid.

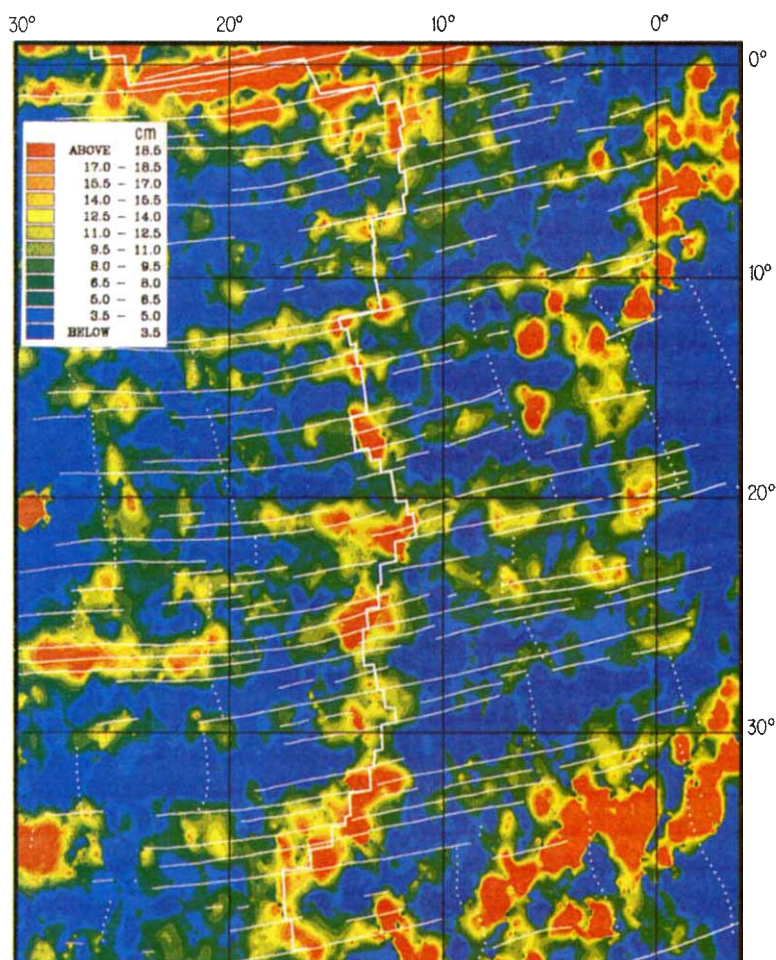
geophysical data set available in this region. Figure 2 shows Seasat tracks crossing the central province of the South Atlantic. They are part of a more extensive data set discussed in detail by Gibert and Courtillot⁹: the along-track sampling interval is 7 km and the smallest noise-free wavelength is ~ 30 km. Spectral analyses of the profiles, and studies of compensation mechanisms, point to a strong correlation between geoid undulations and bathymetry for wavelengths < 300 km (ref. 9). Although compensation mechanisms may vary in nature and efficiency, the 30–85-km waveband corresponds to essentially uncompensated bathymetry and leaves little doubt that short-wavelength geoid undulations are geophysically significant.

The short-wavelength geoid anomalies have been isolated by along-track filtering of the raw Seasat data, using efficient recur-

sive bandpass filters. The cutoff wavelengths were chosen to be 25 and 110 km (3 dB attenuation) and the gain is accurately close to unity for wavelengths between 30 and 85 km (ref. 9). The filtering procedure and orientation of the satellite profiles (Fig. 2) ensures good recovery of all features orientated within 30° of an east-west direction¹⁰.

We have shown⁹ that, with reasonable assumptions about the spectral characteristics of the poorly known bathymetry, geoid undulations with a wavelength < 200 km, say, are representative of bathymetric features overlying a weak lithosphere with an elastic thickness in the range 2–5 km: this has been shown to be true both for the ridge and its flanks between 10° S and 25° S (ref. 9). Given the relationship between elastic thickness and age of the lithosphere¹¹, the bathymetric features seem to have

Fig. 4 Geoid roughness map derived from the Seasat profiles shown in Fig. 2 and processed according to the example in Fig. 3. Contours drawn at 1.5-cm intervals; range from 3.5 cm (deep blue) to 18.5 cm (red). Also shown are the Mid-Atlantic Ridge, the fracture-zone pattern and simplified magnetic chrons (anomalies 13, 29 and 34; J. L. Olivet, personal communication). The most prominent roughness features correspond to the major equatorial fracture zones (One South, Romanche, Chain), to the aseismic ridges (Walvis, Rio Grande, St Helena-Sao Tome), and to the Mid-Atlantic Ridge. Note the quasi-periodic variations of the roughness along the Mid-Atlantic Ridge. The ridge flanks are characterized by intermediate roughness features arranged into elongated ribbons following flow lines and connected to roughness maxima on the Mid-Atlantic Ridge. Mercator projection.



been emplaced on young lithosphere^{9,12}. Short-wavelength geoid undulations are therefore expected to reflect primarily volcanic and tectonic activity fairly close to the ridge.

High-pass filtered maps of the geoid (Fig. 2) can be used to infer the geometry of flow lines and hence to recover details of the kinematic history of the South Atlantic^{10,13,14}. We are interested here in displaying the regional variations of the amplitude of these short-wavelength anomalies. After experimenting with various techniques, including an L1 norm method, we have selected the envelope (that is, the modulus of the analytic signal, related to the Hilbert transform; see ref. 15), which is computationally simple and efficient. The various techniques actually provide very similar results. Figure 3 shows the various computational steps on a typical profile. The envelope (dashed line) of the original high-pass filtered profile (thin solid line) still contains very-short-wavelength components of small amplitude, which are next filtered out by means of a low-pass filter with a cutoff at 180 km (thick upper solid line). This low-pass filtered envelope contains the information we are interested in: it displays the regional variations of the amplitude (or variance) of the short-wavelength geoid, which we will from now on call geoid roughness.

Roughness was calculated for all profiles shown in Fig. 2, and Fig. 4 shows the resulting roughness map, constructed using a bilinear interpolation, and also the Seasat-derived plate boundary¹ and flow lines¹⁰ for comparison and easy localization. This coloured map, part of a larger map comprising the central and South Atlantic from 20° N to 50° S (ref. 10), provides a striking overall view of South Atlantic roughness, where individual features are quickly identified.

The most prominent high-roughness features correspond to the major equatorial fracture zones (One South, Romanche and

Chain) and to the major aseismic ridges (Rio Grande, Walvis and a line going through the St Helena, Sao Tome and Principe Islands), which are generally interpreted as volcanic hotspot traces^{16,17}. The latter line is parallel to the Walvis ridge and far more prominent on the roughness map than on most bathymetric charts, and the southeasternmost termination of the Rio Grande Rise shown in Fig. 4 does not correspond to features found on the GEBCO charts¹⁸; it is a linear extension of the rise mapped further to the north-west and results in an overall feature more symmetrical with respect to the Walvis ridge¹⁰. The next most prominent feature in the roughness map is the Mid-Atlantic Ridge, but roughness is not uniform along strike but rather varies in a quasi-periodical fashion. About eight distinct roughness maxima can be distinguished between 2° S and 32° S, that is, between the Chain and the Tristan da Cunha fracture zones. Five maxima correspond to corners of fracture zones (FZs) with significant offsets: the western corner of the Ascension FZ near 7° S, the two corners of the Bode Verde FZ near 12° S, the eastern corner of an FZ near 25° S and the western corner of another near 32° S. The northernmost maximum is on the ridge and may correspond to an FZ, but one with a small offset; it also centres on a significant crenulation (or sinusoidal wave in the plate boundary). A maximum near 17° S is also at the centre of a crenulation, and may be associated with two FZs with significant (and opposite) offsets. Two maxima are more remote from the plate boundary, and are at the centre of a complex crenulation. In summary, there is an indication that maxima correspond to FZs with a significant offset, but this is not a one-to-one correspondence.

The spacing of the maxima is rather uniform, although it may decrease from north to south. Depending on whether a cluster of three maxima near 12° S is taken as one or two separate

maxima along the ridge, the average spacing is found to be 450 km ($N=7$, variance 75 km) or 400 km ($N=8$, variance 85 km), measured orthogonal to the flow lines. The maxima form a remarkably straight alignment, which actually passes through the more irregular (sinuous) plate boundary.

The ridge flanks are characterized by intermediate roughness. Higher roughness features are arranged in broad elongated anomalies that follow the flow lines and appear to be connected to roughness maxima on the plate boundary (Fig. 4). This indicates that the roughness generating process has a memory over tens of Myr. The ~ 400 -km wavelength features described here on a map of geoid roughness (that is, amplitude modulation of shorter wavelengths; Fig. 4) should not be confused with features of similar wavelength observed in the geoid, for example, in the Pacific and Indian oceans^{19,20}.

Together with modulation of roughness along flow lines, these observations confirm that most of the energy in short-wavelength anomalies of the geoid is related either to transform faults (and fracture zones) or to ridge-related processes (for review, see ref. 21) or to both. The Brazil and Angola basins show as zones of very low roughness. At the boundary between the ridge flanks and the basins (between the anomaly 21 and 29 isochrons, that is, 50–65 Myr BP²², there are symmetrical bands of enhanced roughness that may reflect the kinematic reorganization which is known to have affected the Atlantic bordering plates at that time^{14,23}.

Although short-wavelength undulations of the geoid are directly related to bathymetry, they provide information only on the presence and/or magnitude of the relief but not on the way in which it was created. Roughness variations associated with FZs have a 'tectonic' origin, those corresponding to aseismic ridges signal constructional volcanic features. Both volcanic and tectonic origins can *a priori* be claimed for features now at the plate boundary or previously generated at the ridge. High roughness is notably present in the major South Atlantic FZ¹⁰; the correspondence between some geoid roughness maxima and FZs along the ridge also suggests a primary tectonic origin for them.

An interesting elongated high-roughness feature is found between 20° W and 30° W and 26° S and 28° S, which follows three nearby FZs (Fig. 4). There is no nearby corresponding bathymetric signature in the GEBCO charts, but Gamboa and Rabinowitz²⁴ have demonstrated there the presence of basement ridges and lows with relief differences of $\sim 1,500$ m. These ridges follow flow lines and are called Rio Grande FZs by Gamboa and Rabinowitz (this is not the Rio Grande FZ defined at the ridge axis by the GEBCO charts; see also ref. 1). This region of tectonized, rough relief could produce short-wavelength geoid anomalies of ~ 0.7 m, which is close to the level observed in the roughness map (Fig. 4).

Although the ~ 400 -km wavelength in geoid roughness may be the clearest such feature described so far, it is not the only one to have been suggested in the Atlantic Ocean. Le Douaran and Francheteau²⁵ estimated variations in zero-age depth along the North Atlantic ridge, between 10° N and 50° N. Their smoothed bathymetric profile shows 11 maxima, with an average spacing of 370 km (variance 130 km); these features correspond to a higher density of FZs and to increased topographic variance, that is, roughness. Le Douaran and Francheteau²⁵ found a correlation of the longest-wavelength features with gravity anomalies and trace element and isotope geochemistry of dredged basalts. Hamelin *et al.*²⁶ provided further evidence for such a correlation, using a much larger set of Pb and Sr isotope data, and interpreted their observations in terms of a two-layer mantle with hot-spot-related blob-like injections into the ridge.

More recently, Schilling *et al.*²⁷ and Hanan *et al.*²⁸ reported geochemical and bathymetric variations along the strike of the South Atlantic Ridge, between 2° S and 47° S. They concentrated on five spike-like anomalies in elemental and isotopic (Pb) ratios, which they attributed to pollution of mid-ocean ridge

basalt by nearby hotspots. Closer observation of their Pb isotope ratios and bathymetric data suggests eight or nine anomalies with a characteristic wavelength close to the one we observe in geoid roughness: these observations underline the potential significance of ~ 400 -km segmentation visible over the entire Atlantic Ocean in geoid roughness, bathymetry and some geochemical anomalies.

Large-scale segmentation along the South Atlantic Ridge is superimposed on the smaller-scale segmentation known so far. Its origin can be interpreted in the same way as used for the smaller features (although the interference of the two scales in space and time remains to be studied further). One could first propose that they reflect the original pattern of continental breakup and are related to the breakup process itself. Next one could invoke the thermal contraction crack model², or propose that roughness reflects fabric generated by normal plate-tectonic processes at fracture-zone/ridge intersections. Finally the distribution of roughness maxima might reflect the planform of upper-mantle convection. Segmentation could then be due to gravitational instabilities, as advocated by Whitehead *et al.*⁶, with smooth zones corresponding to excess volcanism, that is, upwellings. The spacing L of upwelling instabilities would depend both on the thickness t of the buoyant layer and on the ratio of the viscosity of the overburden to that of the deep source layer. Recent experiments on the diapirism of cylindrical sources by Kerr and Lister^{29,30} seem to argue for a simpler relation ($L \approx 5t$) with no dependence on the viscosity ratio. If we were to follow the results of the Kerr and Lister experiments, our observed value of large-scale segmentation in the South Atlantic would lead to a source diameter (that is, both width and typical thickness) of ~ 80 km. This does not contradict recent discussions on the extraction of melt from the asthenosphere^{31–33}.

We thank C. J. Allègre, A. Cazenave, L. Fleitout, J. Francheteau, C. Jaupart, K. MacDonald, S. Miller, A. Nicolas, J.-L. Olivet, B. Parsons, J. P. Poirier and A. B. Watts for comments, and Ross Kerr for giving us a copy of his paper (ref. 30) before publication. This work was partly supported by the ATP 'Télédétection Spatiale' of the Centre National de la Recherche Scientifique.

Received 9 November 1987; accepted 7 March 1988.

1. Sandwell, D. T. *J. geophys. Res.* **91**, 6405–6417 (1986).
2. Schouten, H. & White, R. S. *Geology* **8**, 175–179 (1980).
3. Schouten, H., Klitgord, K. D. & Whitehead, J. A. *Nature* **317**, 225–229 (1985).
4. Abbott, D. *Geophys. Res. Lett.* **13**, 157–160 (1986).
5. Turcotte, D. L. *J. geophys. Res.* **79**, 2573–2577 (1974).
6. Whitehead, J. A., Dick, H. J. B. & Schouten, H. *Nature* **312**, 146–148 (1984).
7. MacDonald, K. C., Sempéré, J. C. & Fox, P. J. *J. geophys. Res.* **91**, 10501–10511 (1986).
8. Tapley, B. D., Born, G. H. & Parke, M. E. *J. geophys. Res.* **87**(C5), 3179–3188 (1982).
9. Gibert, D. & Courtillot, V. *J. geophys. Res.* **92**, 6235–6248 (1987).
10. Gibert, D., Courtillot, V. & Olivet, J. L. *J. geophys. Res.* (submitted).
11. Watts, A. B. *J. geophys. Res.* **83**, 5989–6004 (1978).
12. Watts, A. B., Bodine, J. H. & Steckler, M. S. *J. geophys. Res.* **85**, 6369–6376 (1980).
13. Cande, S. C., LaBrecque, J. L., Haxby, W. F., Mello, S. & Rawley, D. B. *Geodynam. Symp., Texas A&M University* (1986).
14. Shaw, P. R. *J. geophys. Res.* **92**, 9363–9375 (1987).
15. Bracewell, R. *The Fourier transform and its applications*, 267–271 (McGraw-Hill, New York, 1978).
16. Duncan, R. A. *Tectonophysics* **74**, 29–41 (1981).
17. Morgan, W. J. *Tectonophysics* **94**, 123–139 (1983).
18. Canadian Hydrographic Service *General Bathymetric Chart of the Oceans (GEBCO)*, 5th edn (Ottawa, 1982).
19. Haxby, W. F. & Weisell, J. K. *J. geophys. Res.* **91**, 3507–3520 (1986).
20. Cazenave, A., Monnerieu, M. & Gibert, D. *Phys. Earth planet. Inter.* **48**, 130–141 (1987).
21. Rona, P. A. *Rev. Geophys.* **25**, 1089–1114 (1987).
22. Kent, D. V. & Gradstein, F. M. in *The Geology of North Atlantic: The Western North Atlantic Region* (eds Tucholke, B. E. & Vogt, P. R.) Vol. M, 45–50 (Geol. Soc. Am., Boulder, Colorado, 1986).
23. Klitgord, K. D. & Schouten, H. in *The Geology of North America: The Western North Atlantic Region* (eds Tucholke, B. E. & Vogt, P. R.) Vol. M, 351–378 (Geol. Soc. Am., Boulder, Colorado, 1986).
24. Gamboa, L. A. & Rabinowitz, P. D. *Earth planet. Sci. Lett.* **52**, 410–418 (1981).
25. Le Douaran, S. & Francheteau, J. *Earth planet. Sci. Lett.* **54**, 29–47 (1981).
26. Hamelin, B., Dupré, B. & Allègre, C. J. *Earth planet. Sci. Lett.* **67**, 340–350 (1984).
27. Schilling, J. G., Thompson, G., Kingsley, R. & Humphris, S. *Nature* **313**, 187–191 (1985).
28. Hanan, B. B., Kingsley, R. H. & Schilling, J. G. *Nature* **322**, 137–144 (1986).
29. Kerr, R. C. & Lister, J. R. *Terra cogn.* **7**, 139 (1987).
30. Kerr, R. C. & Lister, J. R. *Earth planet. Sci. Lett.* (in the press).
31. Nicolas, A. *J. Petrol.* **27**, 999–1022 (1986).
32. McKenzie, D. J. *J. Petrol.* **25**, 713–765 (1984).
33. Marsh, B. D. *Rev. Geophys.* **25**, 1043–1053 (1987).

Supercritical two-phase separation of hydrothermal fluids in the Troodos ophiolite

Jon Cowan & Joe Cann

Department of Geology, The University, Newcastle-upon-Tyne NE1 7RU, UK

The hot (350 °C) black smoker springs of the ocean floor are the vents of hydrothermal systems in which cold seawater descends below the ocean floor, is heated, reacts with the host rock and is transformed into metal-bearing brine. The salinity of these brines varies between vents from ~0.7 to 2.0 times seawater salinity¹⁻⁴; this variation has been ascribed to crustal hydration⁵, crystallization of chloride-bearing minerals⁶ or two-phase separation of seawater into components of differing salinity^{4,7-12}. Here we present fluid inclusion evidence for two-phase separation in the ancient Troodos sulphide-forming systems of Cyprus and discuss, with the NaCl-H₂O system as a model, how this process could explain observed salinity variations in hydrothermal fluids.

The system NaCl-H₂O (ref. 7) differs from that of pure water in several ways. Figure 1 shows a *P-T* diagram for a fluid of seawater salinity (3.2 wt% NaCl). A curve corresponding to the boiling curve for seawater separates a single fluid field from one where two fluids are stable. This two-phase field is bounded by the liquid-solid NaCl field. When the path followed by rising hydrothermal seawater meets the two-phase curve, the seawater, in response to the decrease in pressure, will unmix into two coexisting phases. If this unmixing occurs below the critical point, a vapour will separate from the liquid-like fluid, which then undergoes *subcritical* two-phase separation, or boiling; however, if the unmixing takes place above the critical point then the fluid cannot boil; instead a dense, highly saline brine will separate. We term this *supercritical* two-phase separation, or condensation. As the fluid moves upwards within the two-phase field, unmixing continues. The denser phase will become increasingly more dense and saline, whereas the less-dense phase will become increasingly less dense and its composition will approach that of fresh water. The overall composition of the fluid remains at that of seawater unless one of the two phases is preferentially removed from the system.

The path followed by most black smoker fluids rising adiabatically to vent at 300–350 °C is also shown in Fig. 1. It does not intersect the two-phase field. If two-phase separation is to produce the observed salinity variations then some of the hydrothermal fluid must be heated to greater temperatures close to the heat source. Recent studies^{7,12,13} have demonstrated that supercritical separation is possible for seawater heated in the high-temperature environment around the top of an ocean-ridge magma chamber. In ancient systems, portions of separated fluids may become trapped in the rock as fluid inclusions. To demonstrate two-phase separation, the inclusions must be coeval and of contrasting salinity, and the observed phase ratios, densities, salinities and homogenization temperatures (*T_h*) must be simply explicable by known phase relations in the NaCl-H₂O system. Studies of hydrothermally altered gabbros dredged from oceanic-ridge settings⁹⁻¹² have revealed sets of fluid inclusions of corresponding high and low salinity which are considered to have trapped supercritically separated seawater but, with one possible exception¹¹, direct relations between these examples and mineralizing systems cannot be demonstrated.

In the Troodos ophiolite, complete fossil analogues of modern spreading-centre hydrothermal sulphide-forming systems are exposed¹³⁻¹⁵ in which upflow zones rise from hydrothermal reaction zones characterized by the presence of epidote assemblages and extreme depletions in Cu and Zn¹³. We relate the hydrothermal reaction zones to hydrothermal circulation at the

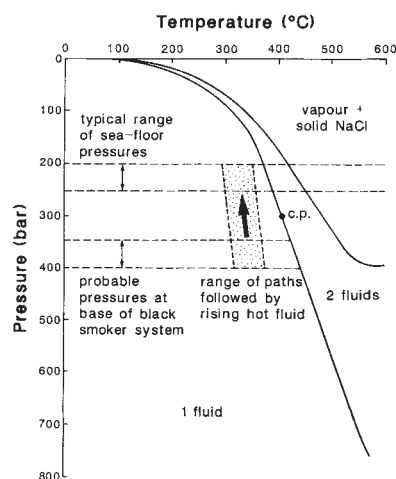


Fig. 1 A *P-T* diagram for an NaCl-H₂O solution of seawater salinity (3.2 wt%) based on data from refs. 23–25. c.p. is the critical point for seawater, which is not, for salt-water systems, the termination of a phase boundary. Seafloor pressures correspond to normal mid-ocean-ridge depths of 2.0–2.5 km. Pressures at the base of the system correspond to a cold hydrostatic pressure 1.5–2.0 km below the seafloor as argued in ref. 13. Black smoker temperatures of 300–350 °C came from refs 1–4 and other sources.

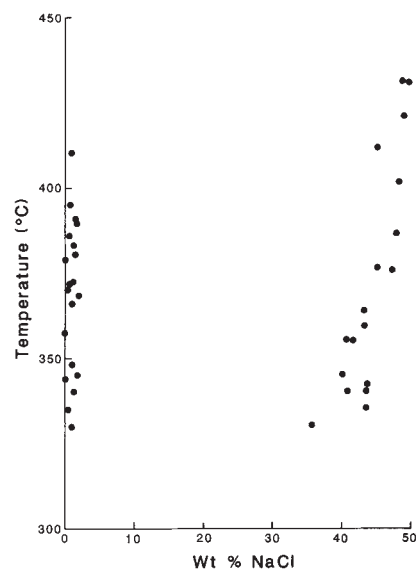


Fig. 2 Diagram of temperatures of homogenization (*T_h*) against salinity for fluid inclusions of sample 6TR 123: compositions were calculated assuming the fluids to be simple NaCl-H₂O solutions. Note the positive correlation between *T_h* and composition for the high-salinity inclusions.

spreading axis for two reasons. First, massive sulphide deposits occur as exhalative lenses within the axial extrusive sequence and thus appear to have formed at the spreading axis. They represent the sink for copper leached from the reaction zones^{13,15}, which therefore must also be a result of on-axis hydrothermal activity. Second, the reaction zones occur just within and immediately above the Upper Plutonic Complex (UPC), which represents the crystallized products of spreading-centre magma chambers¹³. We do not find the arguments for an off-axis origin¹⁵ convincing.

The hydrothermal reaction zones contain many populations of fluid inclusions; those we describe here occur in plagiogranites of the UPC which are considered to represent rock formed at the top of the ancient magma chamber just below the Sheeted

Xenophontos, Ros Strens, Hugh Richards and Dave Manning.
J.G.C. acknowledges receipt of an NERC research studentship.

Received 14 December 1987; accepted 7 March 1988.

1. Michard, G. *et al. Earth planet. Sci. Lett.* **67**, 297-307 (1984).
2. Von Damn, K. L. *et al. Geochim. cosmochim. Acta* **49**, 2197-2220 (1985).
3. Von Damn, K. L., Edmond, J. M., Measures, C. I. & Grant, B. *Geochim. cosmochim. Acta* **49**, 2221-2237 (1985).
4. Von Damn, K. L. & Bischoff, J. L. *J. geophys. Res.* **92**, 11334-11346 (1987).
5. Seyfried, W. E., Berndt, M. E. & Janecky, D. R. *Geochim. cosmochim. Acta* **50**, 469-475 (1986).
6. Vanko, D. A. *Am. Mineral.* **71**, 51-59 (1986).
7. Bischoff, J. L. & Rosenbauer, R. J. *Earth planet. Sci. Lett.* **68**, 172-180 (1984).
8. Stakes, D. & Vanko, D. A. *Earth planet. Sci. Lett.* **79**, 75-92 (1986).
9. Vanko, D. A. *J. Geophys. Res.* (in the press).
10. Kelley, D. S. & Delaney, J. R. *Earth planet. Sci. Lett.* (in the press).
11. Delaney, J. R., Mogk, D. W. & Mottl, M. J. *J. geophys. Res.* **92**, 9175-9192 (1987).
12. Cann, J. R., Strens, M. R. & Rice, A. *Earth planet. Sci. Lett.* **76**, 123-134 (1985).
13. Richardson, C. J., Cann, J. R., Richards, H. G. & Cowan, J. G. *Earth planet. Sci. Lett.* **84**, 243-253 (1987).
14. Richards, H. G. thesis, Univ. Newcastle upon Tyne (1987).
15. Schiffman, P., Smith, B. M., Varga, R. J. & Moores, E. M. *Nature* **325**, 423-425 (1987).
16. Heaton, T. H. E. & Sheppard, S. M. F. in *Volcanic Processes in Ore Genesis*, 42-57 (Geological Society of London, 1977).
17. Roedder, E. *Mineral Soc. Am. Rev. in Mineral.* **12** (Mineralogical Society of America, Washington, 1984).
18. Potter, R. W., Clyne, M. A. & Brown, D. L. *Econ. Geol.* **73**, 284-285 (1978).
19. Keevil, N. B. *Am. chem. Soc. J.* **64**, 841-850 (1942).
20. Spooner, E. T. C. & Bray, C. J. *Nature* **266**, 808-812 (1977).
21. United Kingdom Steam Tables in S. I. Units (Arnold, E., London, 1970).
22. Urusova, M. A. *Russian J. inorg. Chem.* **20**, 1717-1721 (1975).
23. Sourirajan, S. & Kennedy, G. C. *Am. J. Sci.* **260**, 115-242 (1962).
24. Chou, I.-M. *Geochim. cosmochim. Acta* **51**, 1965-1975 (1986).
25. Bischoff, J. L. & Pitzer, K. S. *Earth planet. Sci. Lett.* **75**, 327-338 (1985).
26. Potter, R. W. *J. Research U.S. geol. Surv.* **5**, 603-607 (1977).

Resource compartmentation and the stability of real ecosystems

John C. Moore*[†] & H. William Hunt*

* Natural Resource Ecology Laboratory and
† Departments of Entomology and Statistics,
Colorado State University, Fort Collins, Colorado 80523, USA

Gardner and Ashby¹ and May² challenged the tenet that ecological stability increases with species diversity. But May concluded that if species are arranged in 'blocks', stability restrictions are relaxed allowing for greater diversity than in the absence of blocking. Recent investigations suggest that communities may contain tightly coupled subunits whose numbers may increase with diversity³⁻⁶. McNaughton⁷ suggested that species may be arranged in resource compartments and that, within these compartments, species interaction strength would decline as diversity increased. Pimm⁸ found little evidence of food-web compartmentation within habitats and few examples between habitats but cautioned that binary data sets based on connectedness were not optimal for detecting compartmentation. We provide evidence of resource compartmentation based on structural characteristics of a belowground connectedness web⁹ and on biomass estimates and nitrogen flux rates from its energy flux-web description. Using other food webs, we demonstrate that the relationship between structure and diversity is similar for different resource compartments, suggesting that these compartments behave as entities.

A connectedness food-web description is a graphical representation of feeding relations and is constructed by directing vectors from prey to predators. Three variables used to characterize and study the stability of connectedness webs are (1) number of species S , (2) connectance C , and (3) average interaction strength i . We measured connectance as the proportion of off-diagonal non-zero elements of the community matrix¹¹. Interactions, designated '1' in the community matrix, were determined by feeding, potential competition (two predators feeding on the same prey) and mutualism. Non-interactions were assigned '0'. Interaction strength was measured as the maximum average interaction strength, $i_{\max}$, based on May's² criterion for food-web stability $\{i(SC)^{1/2} < 1\}$.

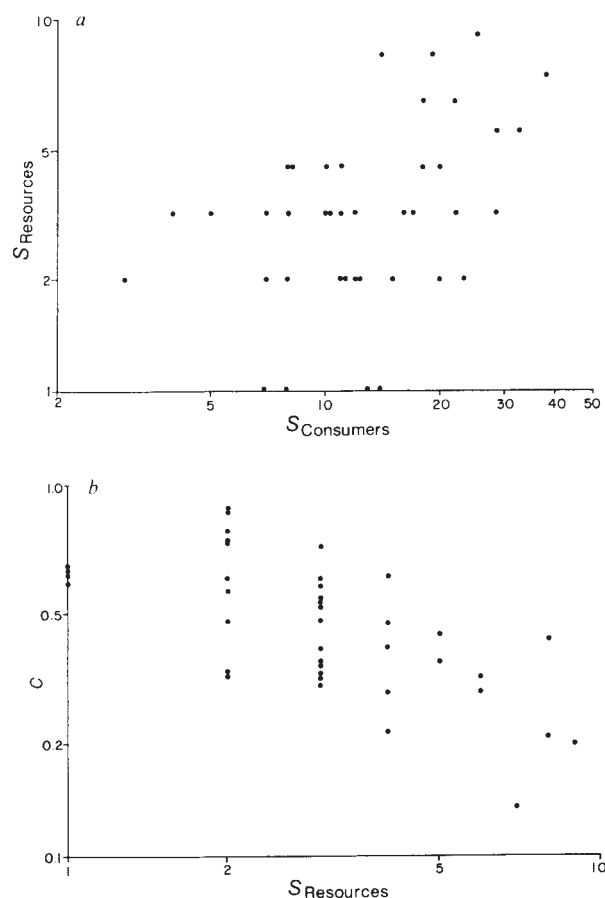


Fig. 1 The relationships between (a) the number of energy channels ($S_{\text{resources}}$) and the number of consumers ($S_{\text{consumers}}$) in the food web ($r = 0.42$, $P = 0.0065$) and (b) the connectance (C) of a food web and the number of energy channels ($S_{\text{resources}}$) in the web ($r = -0.66$, $P < 0.001$).

The belowground connectedness web has 19 'species' aggregations (functional groups and resource types). Functional groups were based on food source, feeding modes, and life-history parameters. The connectance, 0.36, and $i_{\max}$ 0.38, are comparable to those of other food webs with similar diversity^{10,11}. The connectedness web is composed of nearly distinct source webs (*sensu* Cohen¹², Odum¹³ and Marples¹⁴) or 'energy channels', that begin with either detritus (labile and resistant substrates) or plant roots. From the energy-flux food-web description^{9,15} we calculated the proportion of a species' diet derived from each energy source. These results (Table 1) also supported resource compartmentation and further differentiated the detritus channel into a fungal and a bacterial channel.

The pattern of energy flow through the three channels of the belowground food web suggests that functional union occurs even higher in the food web than implied by the connectedness description. Primary decomposers (bacteria and fungi) attack classes of chemical bonds irrespective of the original trophic level or functional group. Thus different energy channels are united by death and decomposition. Higher in the food web, predators have evolved to capture morphologically similar prey that may derive energy from different channels. For example, nematophagous mites prey on bacterial feeding nematodes (bacteria channel), fungal feeding nematodes (fungi channel), and root feeding nematodes (root channel). For this reason the connectedness food-web description appears reticulate at intermediate levels. But considering the energy-flux food-web description, predators at intermediate trophic levels (predaceous nematodes and nematophagous mites) obtained the bulk of their energy from a single resource compartment, whereas it was the

Table 1 The proportion of energy derived from the bacterial, fungal and root energy channels by different faunal groups (including ciliates) in the North American shortgrass steppe

Faunal group	Energy channel Nitrogen flux (mg N m ⁻² yr ⁻¹)		
	Bacteria	Fungi	Roots
Protozoa			
flagellates	100	0	0
amoebae	100	0	0
ciliates	100	0	0
Nematodes			
root-feeders	0	0	100
fungivores	0	90	10
omnivores	100	0	0
bacteriavores	100	0	0
predators	68.67	3.50	27.83
Microarthropods			
mycophagous collembolans	0	90	10
mycophagous oribatid mites	0	90	10
mycophagous prostigmatid mites	0	90	10
nematophagous mites	66.70	3.78	29.52
predaceous mites	39.54	38.56	21.91

This table is based on nitrogen flux rates as calculated in ref. 9. This analysis is similar to that in ref. 15, but differs in that it includes feeding preferences and uses nitrogen flux rates rather than biomass. The root energy channel includes both plant roots and mycorrhizal fungi (VAM). We have assumed that material flows and energy are interchangeable expressions of food-web structure and function. The energy transferred by trophic interactions between web components consists of chemical potential energy associated with covalent bonds. The energy and nitrogen content per unit dry weight of organisms are much less variable than are patterns of flow through webs. Therefore, energy, dry weight and nitrogen flows are equally useful indices of flow patterns through food webs.

top predators (predaceous mites) that strongly linked resource compartments (Table 1).

Pimm⁸ noted that food webs may also be compartmented along the time and habitat niche axes¹⁶. Elliott *et al.*¹⁷ and Moore¹⁸ found that in no-till winter wheat fields in Colorado, consumers within the fungal, bacterial and root energy channels peaked at different periods during the growing season. Hendrix *et al.*¹⁹ and Ingham *et al.*²⁰ provided evidence of this pattern from studies in agroecosystems in the Georgia Piedmont and the shortgrass steppe in Colorado, respectively. Compartmentation in time was also found in aquatic and plant-mutualist food webs^{6,21,22}. On the habitat niche axis, the bacteria channel contains many 'aquatic' organisms; bacteria, protozoa, and nematodes, whereas the fungi channel includes organisms that require high humidity but not free soil water or water films.

The reality of fungal and bacterial resource compartments is supported by their different responses to disturbance. Studies of the dynamics of the belowground food web have indicated that the bacterial energy channel is more resistant to natural and anthropogenic disturbances than the fungal energy channel^{17-19,23,24}. In all cases, organisms within the fungal energy channel responded to disturbance as a unit, whereas bacteria and their consumers showed little or no response.

We characterized 138 energy channels from the 40 community connectedness webs compiled by Briand¹⁰. Although the types of resources at the base of the energy channels were diverse, they could be categorized as either primary producers, detritus or consumers. More than half of the resources (63%) were primary producers: vascular plants, bryophytes and algae. Detritus accounted for 20% of resources, whereas consumers like birds, nematodes and zooplankton, represented 17%. A more comprehensive description of consumer food sources would have allowed these organisms to be placed in either the

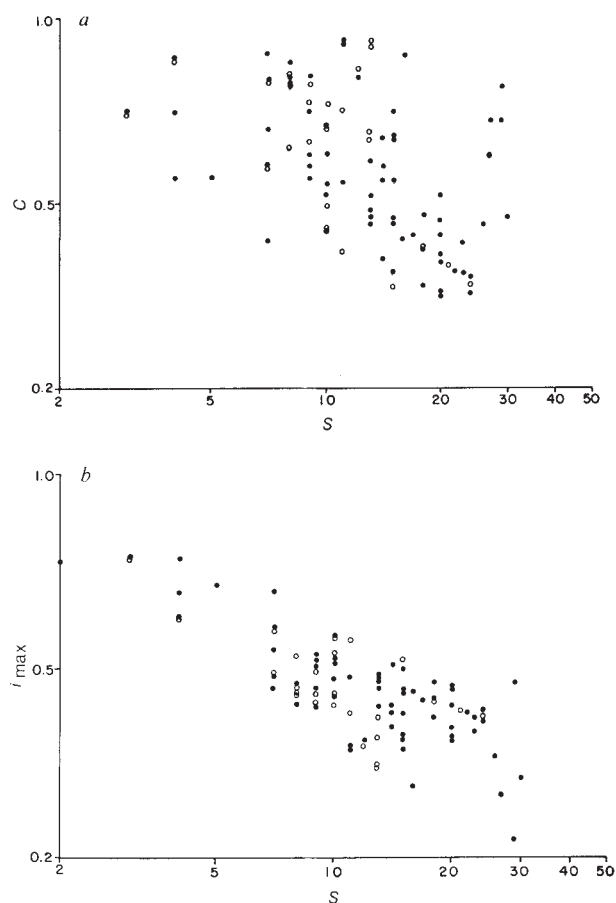


Fig. 2 The relationships between the (a) connectance (C) and the diversity (S) of primary producer (●), ($r = -0.66$, $P < 0.001$) and detritus (○), ($r = -0.57$, $P = 0.0017$) energy channels (combined; $r = -0.64$, $P < 0.001$), and (b) $i_{\max}$ and the diversity (S) of primary producer ($r = -0.86$, $P < 0.001$) and detritus ($r = -0.71$, $P < 0.001$) energy channels (combined; $r = -0.83$, $P < 0.001$). For the relationships in both a and b a simple likelihood ratio test (SLRT) showed that the regression equations for primary producers and detritus did not differ significantly at the $P < 0.10$ level. Duplicate points were not included for clarity. The regressions for the primary producers and the SLRT to compare detritus and primary producer energy channels include more than one data point from a web. For example, there were two or more primary producers in a web or the web contained both detritus and primary producers.

detritus or primary producer energy channels or both. The majority of food webs (72.5%) contained both detritus and primary producer energy channels, whereas the remaining food webs had other configurations. Habitat compartmentation was also evident as 26.8% of the food webs^{9,10} either contained both aquatic and terrestrial based energy channels, or were aquatic based with a large number of terrestrial and aquatic consumers.

Using Briand's webs, we explored how energy channels affect food-web structure and how species diversity within an energy channel is constrained by connectance and $i_{\max}$. The number of energy channels (resource bases) per food web increased with diversity (Fig. 1a), and the connectance of food-web declined with an increased number of energy channels (Fig. 1b). As expected^{25,26}, $i_{\max}$ remained constant (average $i_{\max}$ was 0.39) across food webs, regardless of the number of energy channels they contain. Energy channels tended to have higher C and $i_{\max}$ than the webs from which they were derived and the proportionate difference in C was an increasing linear function of the number of energy channels ($r = 0.72$, $P = 0.001$). But the average connectance within energy channels (0.60) was not a

function of the number of energy channels in the web ($r = 0.08$, $P = 0.63$). These results suggest that (1) species interact more within a given energy channel than with species in other energy channels and (2) criteria governing the structure and stability of an energy channel do not include other energy channels.

McNaughton⁷ predicted that resource compartments would be constrained by the same stability criteria as whole webs; that is, as diversity increased, either the connectance or interaction strength declined. Our results support this prediction and further illustrate that detritus and primary producer based energy channels do not differ in terms of their structure and the relationship between structure and stability (Figs 2a, b).

Energy channels, and habitat and time differences make food webs less constrained by connectance than connectedness descriptions imply and, consistent with niche theory^{27,28}, serve to reduce interactions among species. This conclusion may explain why only 6 of the 41 connectedness food-web descriptions analysed here have S or C values within the stability thresholds previously established¹. Typically, connectedness food webs depict both weak and strong interactions among species aggregations or resource types^{4,15}. Our study supports the view^{4,8,29} that analyses of the utilization of energy coupled with qualitative analyses of connectedness webs are more likely to reveal aspects of food-web structure and stability than the qualitative analyses alone, as analyses that use energy flows attempt to weight and separate the vectors of interactions among species.

We thank Mr Martin Fowler for writing the program to calculate the food-web parameters, Dr P. L. Chapman for his

comments on the statistical analysis and Drs D. C. Coleman, M. Coughenour, E. T. Elliott, D. S. Schimel and D. E. Walter for their comments. This work was supported by grants from the DOE and the NSF.

Received 1 February; accepted 8 April 1988.

1. Gardner, M. R. & Ashby, W. R. *Nature* **228**, 274 (1970).
2. May, R. M. *Nature* **238**, 413-414 (1972).
3. Root, R. B. *Ecol. Monogr.* **43**, 95-124 (1973).
4. Paine, R. T. *J. anim. Ecol.* **49**, 667-686 (1980).
5. Yodzis, P. *Nature* **284**, 544-545 (1980).
6. Jordano, P. *Am. Nat.* **129**, 657-677 (1987).
7. McNaughton, S. J. *Nature* **274**, 251-253 (1978).
8. Pimm, S. L. *Food Webs* (Chapman and Hall, London, 1982).
9. Hunt, H. W. *et al. Biol. Fert. Soil* **3**, 57-68 (1987).
10. Briand, F. *Ecology* **64**, 253-263 (1983).
11. Yodzis, P. *Am. Nat.* **120**, 551-570 (1982).
12. Cohen, J. E. *Food Webs and Niche Space* (Princeton University Press, Princeton, 1978).
13. Odum, E. P. *Jap. J. Ecol.* **12**, 108-118 (1962).
14. Marples, T. G. *Ecology* **47**, 270-277 (1966).
15. Moore, J. C. *et al. Ann. Rev. Ent.* 419-439 (1988).
16. Schoener, T. W. *Science* **185**, 27-39 (1974).
17. Elliott, E. T. *et al. Pl. Soil* **79**, 149-155 (1984).
18. Moore, J. C. Thesis, Colorado State University (1986).
19. Hendrix, P. F. *et al. Bioscience* **36**, 374-380 (1986).
20. Ingham, E. R. *et al. J. appl. Ecol.* **23**, 597-614 (1986).
21. Lane, P. *Ecology* **67**, 223-239 (1986).
22. Kitching, R. L. *Oikos* **48**, 280-288 (1987).
23. Holland, E. A. & Coleman, D. C. *Ecology* **68**, 425-433 (1987).
24. Hunt, H. W. *et al. Pl. Soil* (submitted, 1988).
25. Auerbach, M. J. *Ecological Communities, Conceptual Issues and the Evidence* (Princeton University Press, Princeton, 1984).
26. Rejmanek, M. & Sary, P. *Nature* **280**, 311-313 (1979).
27. Hutchinson, G. E. *Cold Spring Harb. Symp. quant. Biol.* **22**, 415-427 (1958).
28. Hutchinson, G. E. *Am. Nat.* **93**, 145-159 (1959).
29. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton University Press, Princeton, 1973).

Sex pheromone released as an aerosol by the moth *Pyrharcia isabella*

Stuart B. Krasnoff* & Wendell L. Roelofs

Department of Entomology, Cornell University, New York State Agricultural Experiment Station, Geneva, New York 14456, USA

In most moth species the female sex pheromone evaporates from the surface of a gland that is everted and exposed to the air¹⁻³. Considerable attention has been directed towards the pheromone biology of arctiid moths because (1) the female pheromone gland is atypical, consisting of a branched tube or pair of tubes that is invaginated into the haemocoel and opens dorsally between the eighth and ninth abdominal segments⁴⁻⁵, and (2) pheromone release behaviour ('calling') is also atypical, involving rhythmic protrusion and retraction of the abdominal tip⁴. This pulsing action, which alternately exposes and occludes the orifice of the pheromone gland, is thought to result in the expulsion of discrete puffs of pheromone in the vapour phase⁴. We now report that calling females of the arctiid species, *Pyrharcia isabella*, emit a visible stream of liquid droplets. This emission consists solely of the sex pheromone, with no incidental material or inert carrier material present in the aerosol. The quantities of pheromone released in this way are the largest known for a female moth. Also this is the first report of an airborne sex-attractant released as a liquid.

In the laboratory we repeatedly observed the emission of a visible material by calling *P. isabella* females (Fig. 1a). This emission, which appears in a fine continuous stream, originates from the dorsum of the abdominal tip near the orifice of the pheromone gland (Figs 1b, 2). Although the co-occurrence of this visible emission with sexual calling behaviour suggested that the pheromone of this moth is released in particulate form,

we investigated two alternative possibilities: that the emission might be (1) a non-pheromonal material that is only incidentally released at the same time as the pheromone or (2) an inert carrier for the pheromone.

Although the emission readily became windborne (Fig. 1a), it appeared to sink in the still air of a closed container. This suggested a system for collecting the emission by sedimentation. A calling female was positioned in a glass shell vial with the abdominal tip 4.0-5.0 cm above a cover slip covering the bottom of the vial. The emission fell from the abdominal tip, impacted on the cover slip, and left a deposit of liquid droplets measuring 5-20 μm in projected diameter (Fig. 1b-d). No dehiscent scales or fragments of scales, such as those reported as pheromone dispersal or transfer agents in other Lepidoptera^{6,7}, were detected in these collections.

Gas-liquid chromatographic (GLC) analyses showed that the deposits contained copious amounts of compounds previously identified from pheromone gland extracts of *P. isabella*^{5,8}. A primary component, 2-methylheptadecane, was present with two minor components, 2-methylhexadecane and 2-methyloctadecane, also present at 2% and 1% of the major component, respectively⁵. Frequently more than 10 μg of the major component was collected during a calling session of less than one hour with release rates averaging 240 ng min^{-1} ($n = 10$; s.e.m. = 10.0; range = 90-440 ng min^{-1}). This is an order of magnitude greater than any release rate previously recorded for a female moth pheromone^{9,10}, and is only rivalled by male *Achroia grisella* which release up to 10 $\mu\text{g h}^{-1}$ of a two-component pheromone¹¹.

Collecting the visible emission by sedimentation followed by analysis for pheromone content allowed us to test whether the pheromone of *P. isabella* travelled with the visible material or dispersed independently of it as a vapour. In the former case a collection from a live female should give more pheromone from the cover slip than from the walls of the vial. In the latter case more pheromone should be recovered from the walls of the vial (which are closer to the release point and more extensive in surface area) than from the cover slip. In collections from ten calling females, the cover slip gave most of the 2-methyl-

* Present address: Boyce Thompson Institute, Tower Road, Ithaca, New York 14853, USA

Fig. 1 *a*, Female *Pyrharcia isabella* in a wind-tunnel during pheromone release. Air is moving from left to right at $\approx 0.3 \text{ m s}^{-1}$. The undulations indicated by the arrow were produced coincident with the rhythmic pulses of the abdominal tip. The insect is a $\approx 2.5 \text{ cm}$ in length. *b*, Collection of *P. isabella* emission. A calling female was coaxed onto a wooden dowel (3.0 mm in diameter) inserted into a cork stopper covered with aluminium foil. The female was then lowered into a glass shell vial (9.5 cm $\times$ 2.2 cm inside diameter) the bottom of which was covered by a glass cover slip (2.2 cm in diameter). Note emission trailing from the abdominal tip to the bottom of the vial. The curved path of descent taken by the emission is due to a positive photophoretic response²² to the beam of a flashlight used to illuminate the subject. *c*, Deposit of emission on cover slip. *d*, Detail of deposit showing collected droplets. Scale bar, 100 μm . By timing the descent of points of disturbance in the pheromone streams from calling *P. isabella* females, terminal velocities of settling particles in the stream were measured at $1\text{--}3 \text{ mm s}^{-1}$ (at $T = 20^\circ\text{C}$). The estimated Stokes' diameter of spherical droplets of 2-methylheptadecane (density 0.78 g cm^{-3}) falling at this rate would range between 6.5 and 11 μm in diameter¹³. This estimate accords well with measurements made through a light microscope of the droplets in the deposits (5–20 μm).

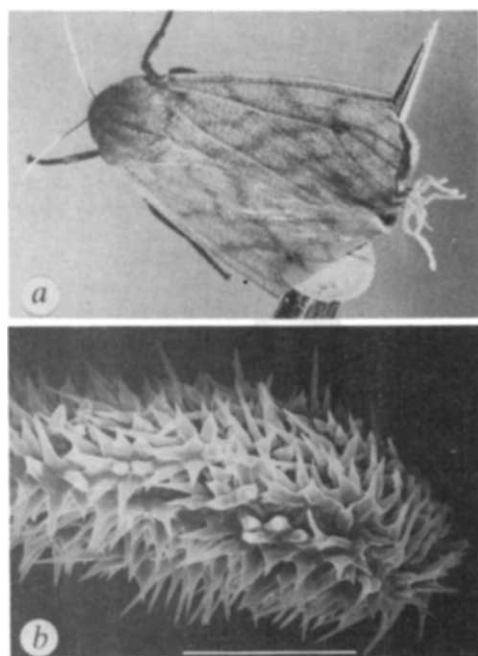
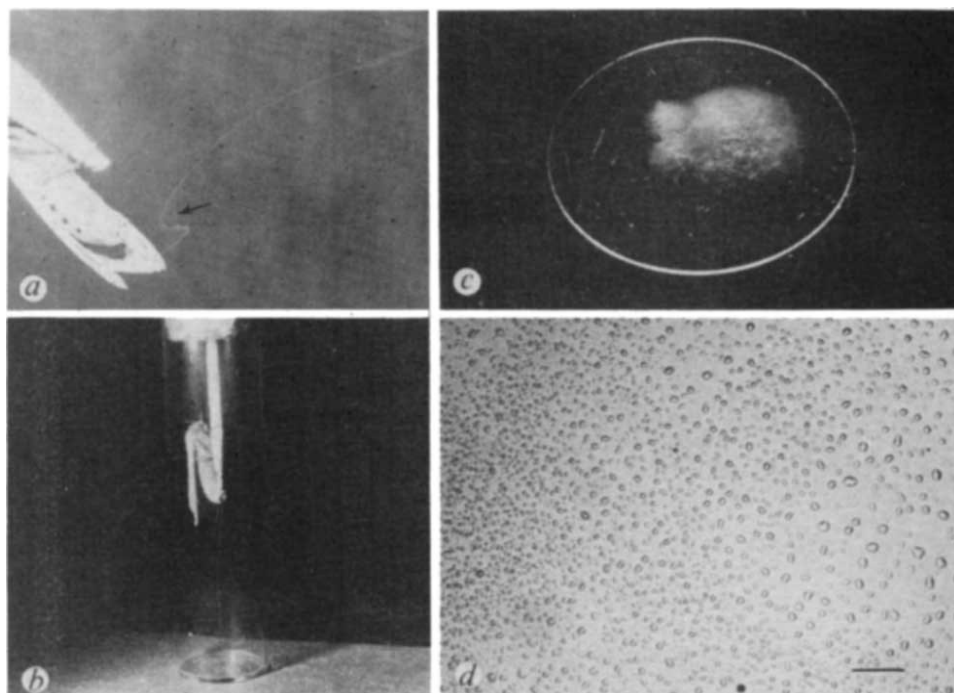


Fig. 2 *a*, Pheromone gland of a *Pyrharcia isabella* female. The gland, which normally lies within the haemocoel, has been artificially everted by squeezing the abdomen to show the morphology of the branches. *b* and *c*, Scanning electron micrographs showing details of the branches of an everted gland. The spines normally project into the lumina of these branches (scale bars, 50 μm in *b*, 20 μm in *c*).

heptadecane (average = 88%; range 75–97%) (Fig. 3*a*). In contrast, collections of 2-methylheptadecane from an evaporating emitter resulted in an average recovery only 5% of the total 2-methylheptadecane from the cover slip (Fig. 3*b*). This is roughly what would be predicted, given that the cover slip had $\sim 1/20$ the surface area of the rest of the vial. This comparison indicated that the pheromone from calling females did not disperse independently of the visible emission, and so the emission could not be composed of a material only coincidentally produced along with the pheromone.

The possibility still remained that the visible emission could be composed primarily of an inert material serving as a carrier for the pheromone. Although pheromonal components accoun-

ted for >95% of the total GLC eluate of the dichloromethane extracts of the emitted deposits, fatty acids and triglycerides, which were considered as possible carriers, would not be directly detectable by the GLC column and temperature programme used. But these fatty acyl compounds were eliminated from consideration because no methyl esters were detected in a GLC analysis of the products of a deposit that had been submitted to acid methanolysis¹². Direct probe mass spectra indicated that the deposits contained a predominance of 2-methylheptadecane with only traces (<5%) of other compounds present. We conclude from these findings that the emission consists solely of the pheromone released as an aerosol¹³ of small droplets.

The release of pheromone as an aerosol may be widespread

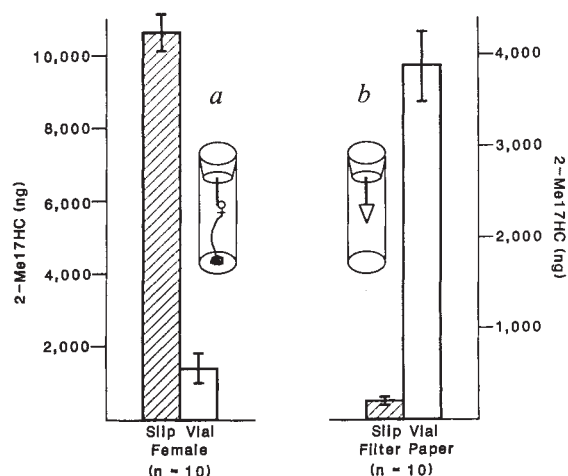


Fig. 3 Mean recoveries ($\pm$ standard error of the mean) of 2-methylheptadecane from cover slip and vial in collections of *a*, emissions from calling *P. isabella* females ($n=10$) and *b*, filter paper treated with 2-methylheptadecane ($n=10$).

Methods. *a*, Collections of *P. isabella* emission were continued until calling ceased (average duration of collection, 45 min) at which time the cover slip was removed and extracted in 1 ml re-distilled dichloromethane, to which an internal standard (hexadecane, 1 μ g) had been added. Extracts stood for at least 1 h at 20 °C and then were analysed on a HP 5840 gas chromatograph with a cross-linked methyl silicone capillary column (internal diameter, 0.31 mm), or a 25-m cross-linked 5% phenylmethyl silicone column (internal diameter 0.25 mm) programmed from 80–230 °C at 15 min⁻¹ after an initial delay of 1 min. Quantification was achieved by comparing electronically integrated area of the 2-methylheptadecane peak with that produced by the internal standard. The interior of the shell vial, the aluminium foil stopper-liner, and the dowel (see Fig. 1*b*) also were rinsed repeatedly with a single 2-ml aliquot of re-distilled dichloromethane. Hexadecane (1 μ g) was added as an internal standard and the rinse stored for >1 h at 20 °C before analysis by GLC as described above. *b*, a filter paper triangle (20 mm $\times$ 20 mm $\times$ 17 mm) was loaded with 20 mg 2-methylheptadecane and placed in a shell vial with its tip 40 mm from the glass cover slip. After 12 h, the vial and cover slip were extracted as in *a*. Correction factors for recovery of 2-methylheptadecane were calculated by extracting cover slips and vials to which known quantities of 2-methylheptadecane had been applied.

in the Arctiidae. We have observed this visible emission in another arctiid species, *Phragmatobia fuliginosa*¹⁴. Also, Schal *et al.*¹⁵ found that another arctiid, *Holomelina lamae*, has a pheromone blend made up of several compounds with varying vapour pressures. The ratio of components is the same in airborne collections from calling females as it is in female gland extracts. It was postulated that this could occur only if the pheromone were released as a liquid rather than as a vapour.

P. isabella and its relatives are the only animals known to release an airborne sex pheromone as a liquid aerosol. Clearly dispersion of such a pheromone takes place as a combination of both the dispersion and the evaporation of the droplets. But the evidence that the droplets may persist in a cohesive stream downwind of a pheromone-releasing female (Fig. 1*a*) suggests that models of pheromone dispersal based solely on molecular diffusion^{2,16,17} are not strictly applicable to these arctiids.

These liquid-phase pheromones also indicate revisions in our thinking about perception of the female pheromone by the male¹⁸. *P. isabella* males appear to have a threshold of response to pheromone that is unusually high for moths, in keeping with the extremely high rate of release of the female pheromone¹⁴.

As noted by others^{9,19}, the geometrid *Rheumaptera hastata* has evolved invaginated tubular pheromone glands independently of the arctiids, and is also reported to pulse the abdomen rhythmically during calling²⁰. Some Lymantriids are also known to pulse¹⁹, and although their glands are not internal tubes, they are deeply folded²¹. Pulsing could have evolved as an effective

mechanism for ejecting pheromone from invaginated or highly convoluted glands in which only a fraction of the total surface area can be exposed to the air. *P. isabella*, with its large extensively branched gland, may represent an extreme expression of a trend toward increased gland surface areas that was accompanied by a revolutionary modification of the form in which the sex pheromone is released.

We thank M. Campbell and K. Poole for assistance in rearing insects, R. Butts for help with mass spectroscopy, and B. Aldwinckle, T. Bourett, W. Meyer, R. Sticht and J. Ogradnick for artwork.

Received 28 January; accepted 18 April 1988.

- Percy, J. & Weatherston, J. in *Pheromones* (ed. Birch, M. C.) 11–34 (North-Holland, Amsterdam, 1974).
- Elkinton, J. S. & Cardé, R. T. in *Chemical Ecology of Insects* (eds Bell, W. J. & Cardé, R. T.) 73–91 (Chapman and Hall, London, 1984).
- Tamaki, Y. in *Comprehensive Insect Physiology, Biochemistry, and Pharmacology* (eds Kerkut, G. A. & Gilbert, L. I.) 145–191 (Pergamon, Oxford, 1985).
- Conner, W. E. *et al. Behav. Ecol. Sociobiol.* 7, 55–63 (1980).
- Meyer, W. L. thesis, Cornell Univ. (1984).
- Leonhardt, B. A., Neal, J. W. Jr, Klun, J. A., Schwarz, M. & Plimmer, J. R. *Science* 219, 314–316 (1983).
- Boppré, M. in *The Biology of Butterflies* (eds Vane-Wright, R. I. & Ackery, P. R.) 259–275 (Academic, London, 1984).
- Roelofs, W. L. & Cardé, R. T. *Science* 171, 684–686 (1971).
- Schal, C. & Cardé, R. T. *Experientia* 41, 1617–1619 (1985).
- Bjostad, L. B., Gaston, L. K. & Shorey, H. H. *J. Insect Physiol.* 26, 493–498 (1980).
- Greenfield, M. D. & Coffelt, J. A. *Behaviour* 84, 287–315 (1983).
- Bjostad, L. B., Wolf, W. A. & Roelofs, W. L. *Insect Biochem.* 11, 73–79 (1981).
- Green, H. L. & Lane, W. R. *Particulate Clouds: Dust, Smokes and Mists* (Spon, London, 1964).
- Krasnoff, S. B. thesis, Cornell Univ. (1987).
- Schal, C., Charlton, R. C. & Cardé, R. T. *J. chem. Ecol.* 13, 1115–1130 (1987).
- Bossert, W. H. & Wilson, E. O. *J. theor. Biol.* 5, 443–469 (1963).
- Mankin, R. W., Vick, K. W., Mayer, M. S., Coffelt, J. A. & Callahan, P. S. *J. chem. Ecol.* 6, 929–950 (1980).
- Roelofs, W. L. *J. chem. Ecol.* 4, 685–699 (1978).
- Conner, W. E. thesis, Cornell Univ. (1979).
- Werner, R. A. *Ann. Ent. Soc. Am.* 70, 264–266 (1977).
- Hollander, A. L., Yin, C. & Schwalbe, C. P. *J. insect. Physiol.* 28, 513–518 (1982).
- Preining, O. in *Aerosol Science* (ed. Davies, C. N.) 111–136 (Academic, London, 1966).

An occlusion-related mechanism of depth perception based on motion and interocular sequence

Shinsuke Shimojo*†, Gerald H. Silverman* & Ken Nakayama*

* The Smith-Kettlewell Eye Research Institute, 2232 Webster St, San Francisco, California, 94115, USA

† Nagoya University School of Medicine, Tsurumai, Showa-ku, Nagoya, 466 Japan

Objects occlude other objects in natural scenes, and this occlusive relationship increases the spatio-temporal complexity of sensory inputs to the two eyes, especially when objects are moving. We ask whether the visual system can employ clever strategies which make use of real-world constraints on inputs to the eyes to determine the depth of objects. Employing psychophysical methods, we found that occlusion-related geometric rules, which constrain the relationship between the direction of motion and the order and asynchrony of eyes, are implemented at early stages of cortical visual processing.

An example of this phenomenon is shown in Fig. 1*a*. When a vertical bar is moved horizontally behind a narrow vertical slit, there is a time lag between the stimuli to the two eyes. At one particular depth, the end of the stimulus to one eye coincides with the onset of the stimulus to the other eye. We call this the 'perfect relay case'. The direction of motion of the bar constrains the sequential order of the monocular stimuli to the eyes: rightward motion is visible to the right eye first, then to the left eye. On the other hand, leftward motion is visible to the left eye first, and then to the right eye. These are the only real-world possibilities if the bar is single and moving laterally behind the

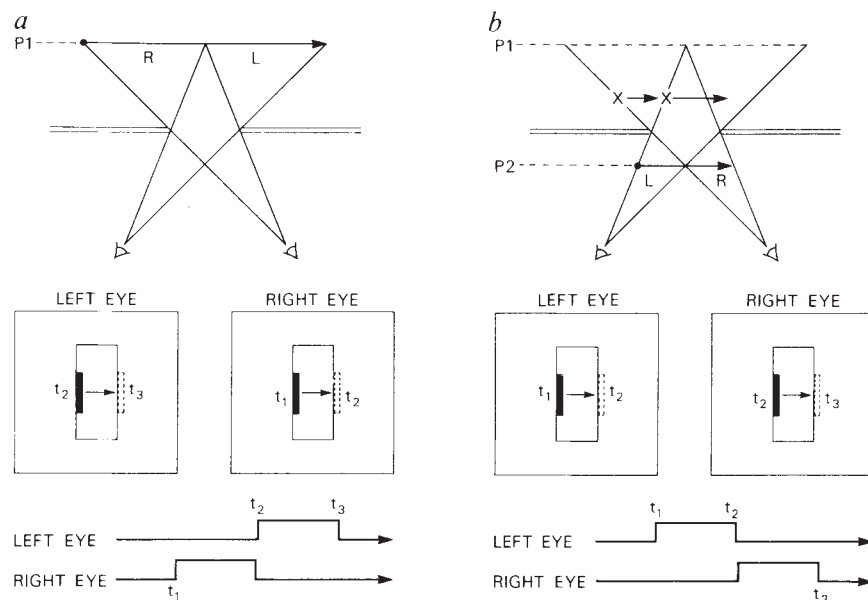


Fig. 1 Schematic diagram of a vertical bar moving behind a vertical slit. The physical structure seen from above (top of figure), the simulated stimulus presented to each eye (middle), and the temporal relationship of monocular stimuli (bottom) are shown in both Fig. 1a and 1b. *a*, The perfect relay case where the target appears in the left eye just as the target in the right eye disappears from view. *b*, The case of reversed ocular order. Smooth motion of a single target at a depth behind the slit was observed in the perfect relay case, whereas two targets moving side by side right behind the slit were observed in the reversed case. In our first experiment, these two stimuli were presented in the top and the bottom halves of the slit. The centre-to-centre separation between the two stimuli was 1.7° , the size of bar target was $32' \times 3.3'$ minute, and the slit width was $20'$ at the observation distance of 80 cm.

slit. Thus, it is theoretically possible to recover its depth relative to the occluder from its direction of motion, the interocular order and asynchrony. Simultaneous binocular disparity is unnecessary in theory.

The question is whether the human visual system actually uses these real-world constraints to determine depth. To study this question, we simulated the perfect relay case on a cathode-ray tube stereo display. Note that, apart from the interocular asynchrony (IOA), each eye receives exactly the same motion trajectory on corresponding retinal points. Our preliminary observations revealed that a seamless motion of a single target is in fact perceived as lying at a depth behind the slit (P1 in the figure). The visual system thus interprets the stimuli which are given sequentially to the two eyes as resulting from one object under smooth motion, rather than from two objects at arbitrary depths.

Reversing the sequence of stimuli to the eyes without changing the direction of motion, however, causes a completely different perception (Fig. 1b). Instead of a single target, the observer sees two bars moving side by side which appear just behind the slit (shown as 'X → X →' at the top of Fig. 1b). Note that if the monocularly viewed target was visually localized on a depth plane in front of the slit, such as P2 in Fig. 1b, it would still violate the geometric constraint posed by occlusion because it should be visible to both the eyes.

To confirm these observations, we tested four subjects (including one naive subject), employing a two-alternative, forced-choice procedure. Two oscillating bars, the perfect relay and the reversed ocular order targets, were presented in the top and the bottom halves of slit in a single test run. Their positions were randomized from trial to trial. The slit width was 20 minutes, and the duration of each monocular stimulus was 50 ms. Thus, the velocity of the targets was 6.7° s^{-1} . The subject's task was to judge which target appeared single, when gazing at a fixation point between them. To avoid possible artifacts related to oculomotor convergence errors, two dots, each presented to one of the two eyes, were added such that when they appeared to align vertically, binocular convergence error would be less than 1.7 min. Subjects were asked to monitor the orientation of these two monocular dots, and to make their judgment only when they were vertically aligned. The percentages of 'single' responses attributed to the perfect relay target were 98% for the naive subject, and 97, 100, and 100% for the other three. These quantitative results agree with our original observations: a target moving far behind the slit is perceived as single only when the direction of motion and the interocular order are in the compatible relationship.

To examine the perceived depth of the slit-motion targets in relation to the conventional stereopsis, a second experiment was carried out. The subject's task here was to adjust the stereoscopic disparity of a comparison target to match the perceived depth of the slit-motion target with various IOAs (Fig. 2). The stereo-matching target was moved sideways in synchrony with the slit-motion target (see inset of Fig. 3) without any occluder so

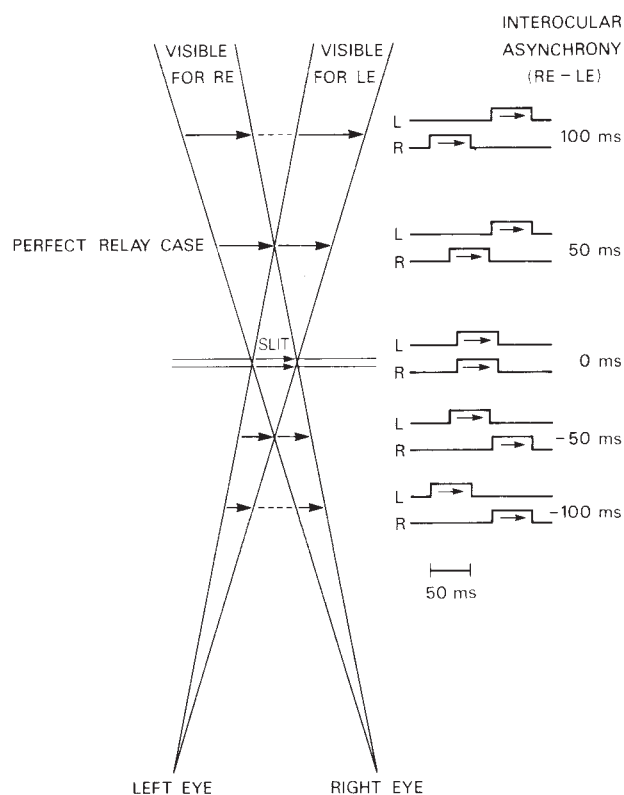


Fig. 2 Relationship between real-world situations of slit view and our simulated motions where an identical target motion is presented within a slit for various degrees of IOA. Note that positive IOAs (right eye followed by left eye when the motion is rightward), including those with no temporal overlap (IOAs greater than the perfect relay case as shown at the top) are compatible with the constraints imposed by occlusion. On the contrary, negative IOAs (left eye followed by right eye) are not compatible with real-world situations because targets moving in front of the occluder should always be seen binocularly. RE, right eye; LE, left eye.

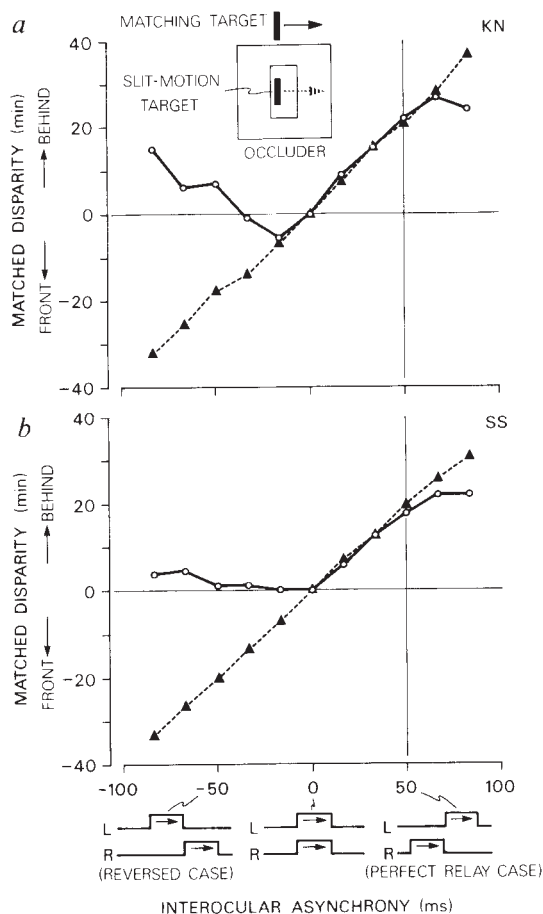


Fig. 3 Perceived depth measured by matching with disparity, as a function of IOA. Two subjects (K.N., top and S.S., bottom) adjusted the disparity of an unoccluded moving target so that it matched the perceived depth of the slit-motion target (open circles). The matching target had a velocity identical to the slit-motion target. Perceived depth of an unoccluded, oscillating stereo target was also measured by the same matching technique (filled triangles). The perfect relay case is marked by a vertical line (IOA = 50 ms). Two findings are noteworthy: first, strong suppressive effects of occlusion cues on negative values of IOA, including the reversed case; second, expanded growth of perceived depth with larger positive values of IOA, even beyond the perfect relay case.

that it was always visible to both the eyes. In the results obtained from two subjects (Fig. 3), two points should be noted. First, the target was not perceived in front of the slit in the reversed case (IOA = -50 ms) or in cases with larger negative IOAs. Second, the perceived depth continued to grow when the IOA increased beyond that of the perfect relay case so that there was an increasing blank interval between the monocular stimuli (50 ms < IOA < 83 ms). The increase in perceived depth with this increasing blank interval cannot be accounted for by the ordinary mechanism of stereopsis which is based on disparities of simultaneous binocular stimulus.

The depth relationship for increasing IOAs beyond the perfect relay case was examined in more detail by our third experiment, in which slit-motion targets with various positive IOAs (0–117 ms) were paired with the standard target of the perfect relay (IOA = 50 ms). The paired stimuli were presented in the top and the bottom halves of slit (see inset of Fig. 4), and the subject was asked to designate the target that appeared farther behind in a forced-choice paradigm. As shown in Fig. 4, results for two subjects confirmed first that perceived depth was larger for the perfect relay case (IOA = 50 ms) in comparison to smaller IOAs where the monocular stimuli overlapped (IOA < 50 ms), and second that perceived depth was clearly larger for IOAs where

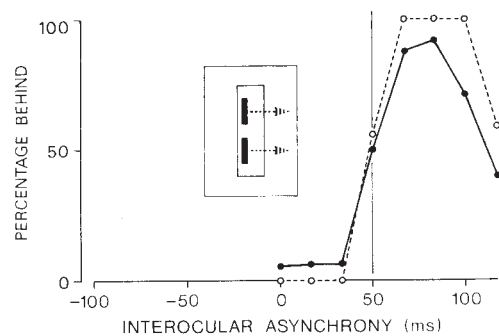


Fig. 4 Relative depth for different IOAs with the perfect relay case as the standard. For both subjects (○, K.N. and ●, S.S.), increasing percentages of the 'behind' response were found for IOAs greater than the perfect relay case (solid vertical line).

there was no overlap and the blank interval was increasing (50 ms < IOA < 100 ms). Thus these and our previous results (Fig. 3) argue for a special mechanism that is not dependent on the processing of fusible disparity.

Our findings taken together cannot be understood in terms of the stereoscopic depth effect which accompanies the interocular delay of a stroboscopically shifting target without temporal overlap of the stimuli to the two eyes^{1–3}. This effect, as well as the 'Pulfrich effect'⁴, can be fully explained by the conventional, binocular stereo mechanisms^{5,6}, which matches the neural effects of monocular spatio-temporal interpolation at a central neural site. These phenomena, unlike our slit-motion phenomenon, are not related to the occlusion constraint, and do not show the asymmetrical effects of disparity reversal which we have seen and which we attribute to occlusive relationships in the real world. The mechanism underlying our slit-motion phenomenon, however, is very different from the mechanisms underlying any of these motion-depth phenomena, or even from the motion in depth detection⁷.

Our findings suggest that the binocular depth system is using cues which are related to occlusion (see refs 8 and 9 for further evidence). This is a somewhat surprising idea because sensory fusion is usually regarded as being mediated rather early, at the striate cortical levels, whereas the encoding of occlusion is often thought to be a late, and even a cognitive, process¹⁰. To consider possible neural mechanisms, note that information regarding direction of motion and interocular order/asynchrony are sufficient to yield appropriate perception of fusion and depth, even when there is no temporal overlap between the inputs to the two eyes. Because eye-of-origin information is critical for determining interocular order and asynchrony, the slit-motion phenomena observed in the current study are unlikely to be mediated at a relatively late cortical site, such as regions VP or MT. This follows from the fact that the majority of neurons in the secondary visual area V2¹¹, in VP¹¹ and in MT¹² can be driven by either eye, and do not appear to retain explicit eye-of-origin information. On the other hand, many neurons in primary visual area V1 are selectively responsive to the eye of stimulation and to the direction of motion¹³. Moreover, some neurons in V1 and V2 are apparently selective to interocular order and asynchrony¹⁴, which may play critical roles in the slit-motion phenomenon.

This research was supported partially by a grant from the National Eye Institute. S.S. is supported by a Fellowship of the Japan Society for the Promotion of Science for Japanese Junior Scientists, and by a Rachel C. Atkinson Fellowship.

Received 29 December 1987; accepted 6 April 1988.

1. Lee, D. *Vision Res.* **10**, 65–78 (1970).
2. Morgan, M. J. *Nature* **256**, 639–640 (1975).
3. Burr, D. C. & Ross, J. *Vision Res.* **19**, 523–532 (1979).
4. Pulfrich, C. *Naturewissenschaften* **10**, 533–564, 569–601, 714–722, 735–743, 751–761 (1922).
5. Julesz, B. *Foundations of Cyclopean Perception* (University of Chicago Press, 1971).

6. Marr, D. & Poggio, T. *Proc. R. Soc. B* **204**, 301–328 (1979).
7. Beverley, K. I. & Regan, D. J. *J. Physiol., Lond.* **235**, 17–29 (1973).
8. Ramachandran, V. S. & Cavanagh, P. *Nature* **317**, 527–530 (1985).
9. Gregory, R. L. & Harris, J. P. *Percept. Psychophys.* **15**, 411–416 (1974).
10. Rock, I. *The Logic of Perception* (MIT Press, Cambridge, 1983).
11. Burkhalter, A. & Van Essen, D. C. *J. Neurosci.* **6**, 2327–2351 (1986).
12. Maunsell, J. H. R. & Van Essen, D. C. *J. Neurophysiol.* **49**, 1148–1167 (1983).
13. Hubel, D. H. & Wiesel, T. N. *J. Physiol., Lond.* **195**, 215–243 (1968).
14. Gardner, J. C., Douglas, R. M. & Cynader, M. S. *Brain Res.* **328**, 154–157 (1985).

Endosomes from kidney collecting tubule cells contain the vasopressin-sensitive water channel

A. S. Verkman, Wayne I. Lencer*, Dennis Brown* & D. A. Ausiello*

Cardiovascular Research Institute, University of California, San Francisco, California 94143, USA

* Renal Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA

The mechanism by which vasopressin rapidly and dramatically increases the water permeability of target epithelial cell membranes is thought to involve a cycle of exo- and endocytosis during which vesicles carrying 'water channels' are successively inserted into, and removed from the apical plasma membrane of epithelial cells^{1,2}. Clusters of intramembranous particles, visible by freeze-fracture electron microscopy and presumed to represent water channels, appear on apical membranes in parallel with increased transepithelial water flow^{3–5}. In the collecting duct, these clusters are located in clathrin-coated pits⁶ which are subsequently internalized. There has been no direct evidence, however, that sub-cellular membranes in vasopressin-sensitive epithelia contain functional water channels. In this report, we have used fluorophores that are sensitive to volume and do not pass through membranes to label and to measure directly the osmotic water permeability of endocytosed vesicles isolated from renal papilla. We present direct evidence that vasopressin induces the appearance of a population of endocytic vesicles whose limiting membranes contain water channels.

To isolate vesicles for analysis of water permeability, homozygous Brattleboro rats with hereditary diabetes insipidus⁷ were dehydrated for 12–22 h and infused intravenously with 6-carboxy-fluorescein (6CF) with and without vasopressin. 6CF is a membrane impermeant fluid-phase marker of endocytosis which filters through the kidney glomerulus, becomes concentrated within the lumen of the collecting tubule and is taken up in collecting tubule cells by endocytosis. A crude microsomal pellet containing these endocytic vesicles with entrapped 6CF was obtained by homogenization of cortical and papillary tissue and differential centrifugation. The osmotic water permeability of 6CF-labelled endocytic vesicles was measured from the time course of decreasing 6CF fluorescence following mixture of microsomal vesicles with hyperosmotic buffer to give a 100 mM inward osmotic gradient (Fig. 1). As water efflux occurs, the concentration of entrapped 6CF increases, resulting in a decrease in fluorescence intensity due to self-quenching. The time course of this process, together with other characteristics detailed later (activation energy, E_a , and water permeability coefficient, P_f), enable a distinction to be made between trans-membrane water movement by lipid diffusion and by bulk flow through water channels. The water permeability characteristics of those sub-cellular vesicles which were not endocytosed during 6CF clearance are not measured in this fluorescence assay.

Vasopressin dramatically alters the kinetics of osmotic water efflux in endosomes from renal papilla (Table 1) which consist almost entirely of endosomes from the vasopressin-sensitive collecting tubule⁸. In endosomes from animals not receiving vasopressin the time course of decreasing fluorescence is fitted

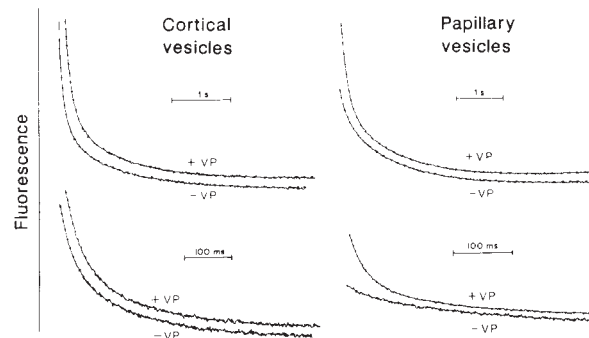


Fig. 1 Time course of osmotic water transport in endocytic vesicles containing 6CF. Brattleboro rats were dehydrated for 12–22 h and infused with a bolus of 5.7 mg 6CF with and without 1.4 μ g vasopressin (Sigma) over 1 min and killed after 15 min. Kidneys were removed immediately and the entire papilla and superficial cortex were dissected in 50 mM mannitol, 5 mM Na phosphate, pH 8.5 (buffer A) at 4°C. Cortical and papillary tissue was homogenized by 15 strokes in a Potter-Elvehjem homogenizer in buffer A at 4°C and centrifuged (500g, 10 min; 5,000g, 10 min). The supernatant was centrifuged at 100,000g for 1 h and maintained in buffer A at 4°C until stopped-flow measurements. Endosomes (~50 μ g protein per ml) were mixed with an equal volume of buffer containing 50 mM mannitol, 200 mM sucrose, 5 mM Na phosphate, pH 8.5 to give a 100 mOsmol per kg H₂O inward sucrose gradient in a Hi-Tech stopped-flow apparatus (dead time <2 ms) interfaced to a MINC/23 computer¹⁸. Fluorescence was excited at 460 nm by a stabilized 100-watt tungsten source in series with a single-grating monochromator and a 450–490-nm six-cavity interference filter. Emission was filtered through two Schott glass OG515 cut-on filters (Duryea). There was no detectable scattered light under these conditions. Each curve is the average of four individual experiments at 21°C. The first 100 ms of the upper curves were expanded and shown in the lower half of the figure; curves were displaced in the y-direction to demonstrate the parallel time courses of the slow exponential processes. In papillary vesicles plus vasopressin (+VP), the drop in fluorescence within the initial 100 ms represents a fast component not present in papillary vesicles without vasopressin (–VP). Biexponential fits (thin beaded lines) are shown through the data; time constants and pre-exponential factors are summarized in Table 1.

by a single exponential function (time constant 1.03 s) indicating the presence of vesicles with a single functional water permeability. In contrast, a bi-exponential function is required to fit the fluorescence curve in endosomes from animals treated with vasopressin, indicating the presence of at least two populations of labelled vesicles. In addition to a population of vesicles with a slow rate of water efflux (1.06 s), a second population of papillary endosomes has a fast rate of water efflux (0.040 s) and an E_a and P_f characteristic of a water channel-mediated process (see below). As expected, vasopressin had no effect on water transport in endosomes isolated from renal cortex, the bulk of which are found in vasopressin-insensitive membranes from proximal tubule⁹. Those cortical endosomes with rapid water transport result from constitutive (vasopressin-independent) endocytosis of water channels known to be present in proximal tubule membranes^{10,11}. These data indicate that vasopressin induces, in the renal papilla, a cellular process resulting in the appearance of 6CF-labelled endosomes whose limiting membranes contain water channels.

We performed control experiments to demonstrate that in our assay the appearance of a water channel in papillary endosomes was a direct effect of vasopressin and not an indirect effect of urine osmolality. 6CF concentration, vascular changes or endosome geometry. Brattleboro rats dehydrated for 12–22 h underwent ureteral catheterization for measurement of urine osmolality and 6CF concentration. In the dehydrated rat, which concentrates its urine in a vasopressin-independent way¹², vasopressin had no further effect on urine osmolality (+vasopressin 598 ± 9 mOsmol/kg, $n=4$) or 6CF concentration (+vasopressin 5.6 ± 1.1 mOsmol/kg, $n=4$) or 6CF

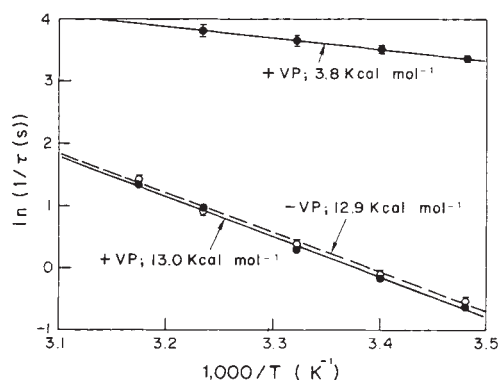


Fig. 2 Arrhenius plot for osmotic water transport in papillary endocytic vesicles. Measurements were performed as described in the legend to Fig. 1 at varying temperatures. Each point is the mean $\pm$ s.d. for measurements performed in quadruplicate. Data were fitted to single activation energies indicated.

concentration (+vasopressin 34 ± 8 mM, -vasopressin 32 ± 2 mM) but still induced the appearance of endosomes containing fast water channels. To show that the vasopressin effect was unrelated to vascular changes, the specific V_2 agonist, 1-D-amino-8-D-Arg-vasopressin (DDAVP, $5\text{--}60 \mu\text{U min}^{-1}$ per 100 g wt) was infused into 18–22 h dehydrated Brattleboro rats. A rapid component of water transport was induced by DDAVP in papillary endosomes, similar to results for vasopressin; there was no effect on water transport in cortical endosomes.

To determine endosome size, we injected rats with 6 mg horseradish peroxidase with and without vasopressin. Microsomal fractions were fixed, reacted with diaminobenzidine and H_2O_2 , and visualized by thin section electron microscopy. Vasopressin had no effect on the size distribution of horseradish peroxidase-positive vesicles which showed a unimodal gaussian distribution of vesicle diameter, with a mean of 110 ± 25 nm. To show that 6CF entered a population of vesicles only by endocytic uptake, 12 mg fluorescein-dextran (relative molecular mass 4,000) was infused instead of 6CF to label endosomes with a fluorophore that had a high molecular mass and did not pass through membranes. Rates of water efflux and the vasopressin effect were identical to those obtained in endosomes labelled with 6CF.

To show that the fast, vasopressin-inducible component of water efflux from endosomes is due to the presence of a water channel, the E_a and P_f for endosomal water transport were determined. In papillary endosomes (Fig. 2) E_a was 13 kcal mol^{-1} for the slow component in vasopressin-treated and untreated animals. E_a for the fast component, present only in the vasopressin-treated animals, was 3.8 kcal mol^{-1} . $E_a > 10$ kcal mol^{-1} is associated with diffusive water movement, whereas $E_a < 5$ kcal mol^{-1} implicates water movement via a channel¹³. For the fast component of water transport in papillary endosomes from vasopressin treated animals, calculated P_f was 0.03 cm s^{-1} at 21°C , based on a 110-nm vesicle diameter. This value is higher than P_f in lipid bilayers ($0.0001\text{--}0.005 \text{ cm s}^{-1}$)¹⁴ and comparable to P_f in red cells (0.02 cm s^{-1})¹⁵ and in the vasopressin-stimulated renal collecting duct ($0.01\text{--}0.02 \text{ cm s}^{-1}$)^{16,17}. The high P_f and low E_a for the fast component of water transport in papillary endosomes isolated from vasopressin treated animals indicate bulk water movement through channels.

These experiments provide direct evidence for a vasopressin-sensitive population of subcellular endocytic vesicles which contain water channels. The ability of vasopressin to induce the appearance of these vesicles in the endocytic compartment of responsive tissue provides strong support for the hypothesis that vasopressin initiates the cycling of functional water channels between the apical membrane and the endosomal compartment. The continuous recycling of water channels allowed us to identify them functionally in the endosomal compartment. The

Table 1 Osmotic water transport in endocytic vesicles

	Time constants (s)		Fractional fast component
	fast process	slow process	
Cortical endosomes			
-vasopressin	0.042 ± 0.002	0.76 ± 0.04	0.62 ± 0.02
+vasopressin	0.042 ± 0.002	0.80 ± 0.05	0.61 ± 0.03
Papillary endosomes			
-vasopressin	Not present	1.03 ± 0.07	0
+vasopressin	0.040 ± 0.004	1.06 ± 0.08	0.44 ± 0.03

Results are mean $\pm$ s.e. for measurements performed on six separate preparations of cortical and papillary endosomes from Brattleboro rats. For each sample, water permeability was measured in quadruplicate and fitted to biexponential functions as in Fig. 1. The fractional fast component is the ratio of the pre-exponential factor for the fast time component to the total signal amplitude.

recycling of these channels would account for the regulatory effect of vasopressin on apical plasma membrane water permeability in target epithelia.

Supported by grants from NIH, the American Heart Association and the Cystic Fibrosis Foundation. A.S.V. and D.B. are established investigators of the American Heart Association.

Received 9 March; accepted 15 April 1988.

- Wade, J. B., Stetson, D. L. & Lewis, S. A. *Ann. N.Y. Acad. Sci.* **372**, 106–117 (1981).
- Hays, R. *Am. J. Physiol.* **245**, C289–C296 (1983).
- Harmanci, M. C., Stern, P., Kachadorian, W. A., Valtin, H. & DiScala, V. A. *Am. J. Physiol.* **239**, F560–F564 (1980).
- Chevalier, J., Bourguet, J. & Hugon, J. S. *Cell Tissue Res.* **152**, 129–140 (1974).
- Kachadorian, W. A., Wade, J. B. & DiScala, V. A. *Science* **190**, 67–69 (1975).
- Brown, D. & Orci, L. *Nature* **302**, 253–255 (1983).
- Valtin, H. *Ann. N.Y. Acad. Sci.* **394**, 1–9 (1982).
- Brown, D., Weyer, P. & Orci, L. *Eur. J. Cell Biol.* (in the press).
- Kriz, W. & Kaissling, B. in *The Kidney: Physiology and Pathophysiology* (eds Seldin, D. W. & Giebisch, G.) 265–306 (Raven, New York, 1985).
- Meyer, M. M. & Verkman, A. S. *J. Membrane Biol.* **96**, 107–119 (1987).
- Verkman, A. S. & Wong, K. R. *Biophys. J.* **51**, 717–723 (1987).
- Valtin, H. & Edwards, B. R. *Kidney Int.* **31**, 634–640 (1987).
- Finkelstein, A. *Distinguished Lecture Series of the Society of General Physiologists* Vol. 4 (Wiley, New York, 1987).
- Fettiplace, R. & Haydon, D. A. *Physiol. Rev.* **60**, 510–550 (1980).
- Macey, R. I. *Am. J. Physiol.* **246**, C195–C203 (1984).
- Hebert, S. C. & Andreoli, T. E. *Kidney Int.* **22**, 526–535 (1982).
- Kuwahara, M., Berry, C. A. & Verkman, A. S. *Kidney Int.* **33**, 435 (1988).

Complete structure of the glycosyl phosphatidylinositol membrane anchor of rat brain Thy-1 glycoprotein

Steve W. Homans, Michael A. J. Ferguson, Raymond A. Dwek, Thomas W. Rademacher, Rita Anand* & Alan F. Williams*

Glycobiology Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

* MRC Cellular Immunology Unit, Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

Glycosyl-phosphatidylinositol (GPI) anchors have recently been identified as alternatives to hydrophobic amino acid sequences for the attachment of a variety of eukaryotic cell surface molecules to the lipid bilayer^{1–4}. In single cell eukaryotes the GPI group appears to be the predominant form of membrane attachment, and in vertebrates a substantial minority of molecules have this anchor including cell surface hydrolytic enzymes, antigens and cell adhesion molecules^{5,6}. Analysis of different GPI anchors suggests they share common structural features including linkage to the COOH group of the terminal amino acid via ethanolamine phosphate, the presence of phosphatidylinositol lipid and a glycan

Table 1 Methylation analysis of the Thy-1 GPI anchor glycan

PMAA	Origin	Neutral glycan	Post JB β HexNAc
Mannitol			
1,5-di-acetyl-2,3,4,6-tetramethyl	terminal Man (Man-4)	1.0	1.0
	(Man-3)		
1,2,5-tri-acetyl-3,4,6-trimethyl	2-0-sub Man (Man-3)	1.7	1.8
	(Man-2)		
1,5,6-tri-acetyl-2,3,4-trimethyl	6-0-sub Man (Man-1)	0.3	1.0
1,4,5,6-tetra-acetyl-2,3-di-methyl	4,6-di-0-sub Man (Man-1)	0.7	0.0
2-deoxy-2-N-methyl-acetamido-galactitol			
1,5-di-acetyl-3,4,6-trimethyl	terminal GalNAc (G)	1.0	0.0

Relative ratios of partially methylated alditol acetates (PMAAs) are shown for the HF treated glycan before and after digestion with Jackbean β N-acetylhexosaminidase (JB β HexNAc). After the hexosaminidase digestion a quantitative conversion of the 4,6-di-O-substituted mannose to 6-O-substituted mannose was observed. From this it is concluded that GalNAc was originally linked to the 4-position of this branch point mannose (Man-1). The presence of 30% of Man-1 as 6-O-substituted mannose in the untreated sample is attributed to partial loss of GalNAc during HF treatment, as no evidence was found for heterogeneity in this region of the molecule from NMR measurements on the unmodified glycan (Fig. 1). The Thy-1 anchor was deaminated with nitrous acid and reduced with NaBD₄ as described⁷, to remove the inositol phosphate group and convert the glucosamine residue to 2,5-anhydromannitol. After desalting of the products⁷ the ethanolamine phosphate groups were removed by treatment with 100 μ l 50% aqueous HF at 0° for 90 h. Following dephosphorylation the HF was removed by precipitation with LiOH and the supernatant was made saturated with respect to NaHCO₃. Acetic anhydride was added (three aliquots of 10 μ l at 10 min intervals) to re-N-acetylate the galactosamine residue (some de-N-acetylation of hexosamines occurs during HF treatment, unpublished data). The resulting neutral glycans were desalted by passage through a tandem column of Chelex 100 (Na⁺)/Dowex AG50X12 (H⁺)/Dowex AG3X4 (OH⁻)/QAE-Sephadex A25 equilibrated and eluted with water. A sample of this neutral glycan fraction (20 nmol) was taken directly for methylation analysis. Another sample (30 nmol) was digested with jackbean β N-acetylhexosaminidase (1.4 units of enzyme in 100 μ l 0.2 M citrate/phosphate buffer pH 5.0, 24 h, 37°). After digestion the pH was adjusted to 10.5 by the addition of 100 μ l 1M NH₄OH and the released GalNAc reduced with 75 μ l 1M NaBD₄ to N-acetylglucosaminitol to prevent its misidentification as a non-reducing terminal GalNAc in the subsequent methylation analysis. Details of the methylation protocols and analysis of the resulting partially methylated alditol acetates (PMAAs) by gas chromatography-mass spectrometry are given in ref. 7.

Table 2 Bio Gel P4 analysis of the GPI neutral glycan fractions

Substrate	Treatment sequence			Major Products	Yield	Size (glucose units)
	aqHF	JB α M	aqHF			
$ \begin{array}{c} \text{Cys} \\ \\ \text{EtN} \\ \\ \text{P} \\ \\ \pm \text{M4-M3-M2} \\ \\ \text{G-M1-A}^* \\ \\ \text{P} \\ \\ \text{EtN} \end{array} $	+	-	-	$ \begin{array}{c} \text{M4-M3-M2} \\ \\ \text{G-M1-A}^* \end{array} $	60%	6.5
				$ \begin{array}{c} \text{M3-M2} \\ \\ \text{G-M1-A}^* \end{array} $	25%	5.7
	+	+	-	G-M1-A*	80%	4.3
	-	+	+	$ \begin{array}{c} \text{M3-M2} \\ \\ \text{G-M1-A}^* \end{array} $	85%	5.7

A portion (40 nmol) of Thy-1 anchor material was deaminated and reduced with NaB³H₄ (12 Ci mmol⁻¹) as described⁷. The products, bearing ³H-labelled terminal 2,5-anhydromannitol residues were purified from radiochemical contaminants by descending paper chromatography⁷. The products, shown under substrate, were treated in three different ways and the neutral products were analysed by Bio Gel P4 chromatography. The size of these products is quoted in glucose units by interpolation of the radioactive glycan elution positions between the positions of co-injected dextran oligomers⁷. Jackbean α -mannosidase (JB α M) digestions were performed at 300 μ M substrate concentration in 0.1 M sodium acetate pH 5.0 containing 100 units ml⁻¹ enzyme, 37°, 24 h. Aqueous HF dephosphorylation re-N-acetylation and de-salting were as described in Table 1 except that shorter (36 h) aqueous HF treatments were used. Under these conditions less hydrolysis of the GalNAc β 1-4Man linkage was observed as indicated by the yields of the GalNAc-containing neutral glycans. The balance of the radioactivity was found in minor peaks corresponding to the structures shown in the table but lacking the N-acetylglucosamine residue, which elute about 1.5 glucose units smaller than those shown in the table. Treatment of dephosphorylated glycans with JB α M resulted in the removal of the M4, M3 and M2 α -mannose residues as expected. Treatment of the phosphorylated glycan with JB α M removed only the M4 α -mannose residue indicating that the M3 residue was originally substituted by the bridging ethanolamine phosphate group, rendering it resistant to JB α M exoglycosidase action. With these data, and despite the lack of firm assignments for the H6 protons of M3, the substitution was assigned to the 6 position from the lack of heteronuclear couplings to the H1, H2, H3 and H4 protons. The residue descriptors are as shown in Fig. 1 except for A*, 2,5-anhydromannitol (³H-labelled); EtN, ethanolamine; P, phosphate.

between the bridging ethanolamine phosphate and the lipid^{5,6}. In the case of the *Trypanosoma brucei* variant surface glycoprotein (VSG) the full structure of the GPI anchor has been determined⁷ and this provides a prototype for comparison with other molecules. We now report the structure of the GPI anchor of rat brain Thy-1 glycoprotein. It has an identical backbone to the VSG anchor but shows significant differences in side chain moieties.

The Thy-1 anchor, attached to the COOH-terminal Cys amino acid, was purified after release of its di (acyl/alkyl) glycerol

groups with *Staphylococcus aureus* phosphatidylinositol-specific phospholipase C (kindly provided by Dr M. Low), as described in the legend to Fig. 1. From 1.2 μ moles of glycoprotein about 0.55 μ moles of anchor was recovered and subjected to analysis by one- and two-dimensional proton NMR (Fig. 1), gas chromatography-mass spectroscopy (GC-MS) (Table 1) and exoglycosidase digestions (Tables 1 and 2). From this analysis the complete structure as shown in Fig. 2a was determined with the exception that it was not possible to distinguish between

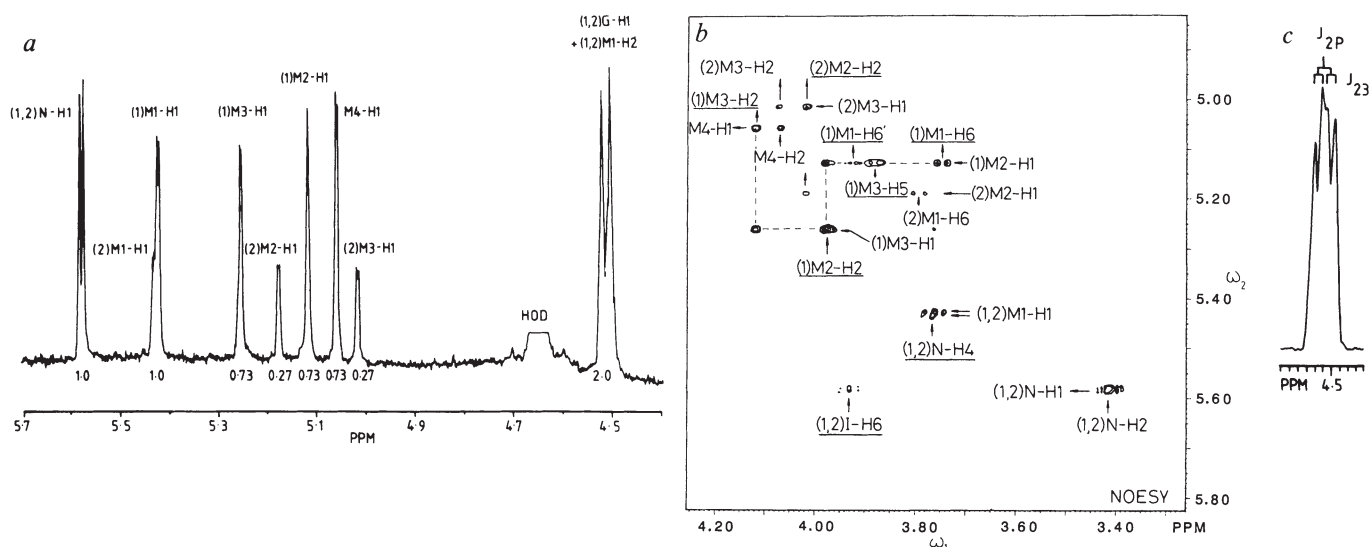


Fig. 1 ^1H NMR spectroscopy of the delipidated Thy-1 GPI anchor. **a**, Anomeric proton region of the 500 MHz ^1H NMR spectrum. Resonances are from *N*-acetylgalactosamine (G), mannoses as in Fig. 2 (M1-4) or glucosamine (N). H1 and H2 indicate the anomeric and H2 proton resonances. The numbers in brackets indicate resonances from the structure with Man-4 (1) or that without Man-4 (2) or unresolved resonances from both structures (1,2). The assignments are based on: anchor composition, spin coupling constants¹⁸, ^1H - ^1H correlated spectroscopy (ref. 19 and refs therein) (COSY, data not shown) and ^1H - ^1H Nuclear Overhauser Effect spectroscopy (ref. 19 and refs therein) (NOESY, see below). Integrated relative intensities are shown under each resonance. **b**, Reporter region of the NOESY spectrum with abbreviations as above plus further proton assignments H 3, 4, 5, 6, 6'. Both intrasidue and inter-residue (underlined) through-space connectivities are observed. For example, the structure $\text{Man}\alpha 1\text{-}2\text{Man}\alpha 1\text{-}2\text{Man}\alpha 1\text{-}6\text{Man}\alpha$ is suggested by the sequential connectivity illustrated by the dotted line. Inositol-1-phosphate at the reducing terminus of the glycan was deduced from the characteristic resonance positions of the H2 and H6 protons of this residue (4.21 p.p.m. and 3.93 p.p.m.)⁷ in the COSY spectrum (not shown). Glucosamine was assigned at the 6 position of inositol from a through-space connectivity between glucosamine H1 and inositol H6 (bottom left)⁷. The linkage of Man-1 to the 4-position of glucosamine was based on an NOE connectivity between Man-1 H1 and glucosamine H4 (ref. 7) (lower centre). *N*-acetylgalactosamine was assigned to the 4-position of Man-1 by GC-MS methylation analysis (Table 1). Composition indicated two ethanolamine residues per anchor². One was assigned to a phosphodiester linkage at the 2-position of Man-1 in all of the glycans based on the anomalous downfield shifted resonance position of Man-1 H2 (4.51 p.p.m.) together with the presence of a heteronuclear splitting (Fig. 1c). Additionally the resonance position of Man-1 H2 shifted upfield to around 4.05 ppm upon cold aqueous HF dephosphorylation (data not shown). The location of the second ethanolamine residue was established to be on Man-3 by α -mannosidase digestion before and after aqueous HF dephosphorylation (Table 2). **c**, Section through the COSY spectrum at $\omega_2 = 4.05$ p.p.m. showing splitting of Man-1 H2 by Man-1 H3 (4.2 ± 2 Hz) and by phosphorus (8.0 ± 2 Hz). The former is larger than that in α -D-mannose¹⁹ due to the electronegativity of the phosphorus substitution. The small coupling between Man-1 H1 and Man-1 H2 (2.0 Hz) is not observed due to low digital resolution.

Methods. Rat brain Thy-1 (1.2 μmol) was purified, ^{14}C S-carboxymethylated and cleaved (40 h at 37 $^\circ\text{C}$) in 20 ml of 0.1 M NH_4HCO_3 with 2% w/w trypsin^{2,20}. The digestion was stopped with phenyl methyl sulphonyl fluoride, lyophilized and suspended in 4 ml of 0.05 M Tris acetate pH 7.4, 0.02% NaN_3 . An equal volume of *S. aureus* PI-PLC (ref. 21) at 30 $\mu\text{g ml}^{-1}$ in the same buffer was added at 0 and 17 h with incubation for 40 h at 37 $^\circ\text{C}$. The digest was extracted three times with two volumes of hexane and the aqueous phase run on a HPLC C18 reverse phase column in 0.01 M NH_4 acetate, 1.5% acetonitrile. The delipidated GPI anchor passed through the column while the other tryptic peptides were bound and final purification was achieved on a BioGel P-30 column in 0.1 M NH_4HCO_3 . Amino acid analysis showed the expected constituents². For NMR the pure anchor fraction was: lyophilized from 0.1 M NH_4HCO_3 ; dissolved in H_2O and passed through a column of Chelex 100 (Na^+) over Dowex Ag-50X12 (H^+) to remove paramagnetic ions; flash-evaporated three times from 99.96% D_2O ; and dissolved in 0.5 ml 99.96% D_2O at ~ 1 mM. NMR spectra were recorded essentially as described⁷.

the ethanolamine attached to the protein (the bridging ethanolamine) and that with the free NH_2 group. The bridging ethanolamine is depicted in Fig. 2a on the basis of comparison with the VSG structure in Fig. 2b which contains only one ethanolamine of known linkage⁷. The full rationale for the structural assignments in the Thy-1 anchor is given in the legends to Fig. 1 and Tables 1 and 2.

The structure in Fig. 2a represents the first complete determination of the glycan of a mammalian GPI and this can be compared with that for *T. brucei* VSG in Fig. 2b. Allowing for the assignment of the bridging ethanolamine, the backbone from protein through to the inositol-1 phosphate of the phosphatidylinositol group is identical in residues and linkages for Thy-1 and VSG. This strongly argues for a common biosynthetic pathway for GPI anchors throughout eukaryotic evolution. In contrast the side chain groups show considerable variation. Notable differences are the attachment of β -*N*-acetylgalactosamine and ethanolamine phosphate in rat brain Thy-1 at Man-1 compared with 2 to 4 α -galactose residues attached to the equivalent man-

nose in VSG structures. The extra ethanolamine phosphate may be a common feature of GPI anchors from species other than single cell eukaryotes since ethanolamine has been detected at levels of about 2 mol mol⁻¹, or greater, in both rat brain and thymus Thy-1 (ref. 2), bovine and human acetylcholinesterase^{4,8}, human decay accelerating factor (DAF, ref. 9), hamster brain Scrapie Prion¹⁰, human placental alkaline phosphatase¹¹ and squid Sgp2 glycoprotein¹². The extra ethanolamine has a free NH_2 group as detected by reductive methylation in Thy-1 (ref. 13), acetylcholinesterase⁴ and DAF (ref. 9). The *N*-acetylgalactosamine is found on all GPI anchors of rat brain Thy-1 but only on about one-third of the groups from thymocyte Thy-1 on the basis of compositional analysis². This residue has thus far not been found in other anchors whose composition is known.

The Man-4 residue in Fig. 2a is seen on 71% of rat Thy-1 GPI anchors on the basis of Bio Gel P4 chromatography (Table 2). This is in accord with a compositional analysis of 2.7 mol mol⁻¹ for the brain Thy-1 anchor² (Man-3 is not detected by

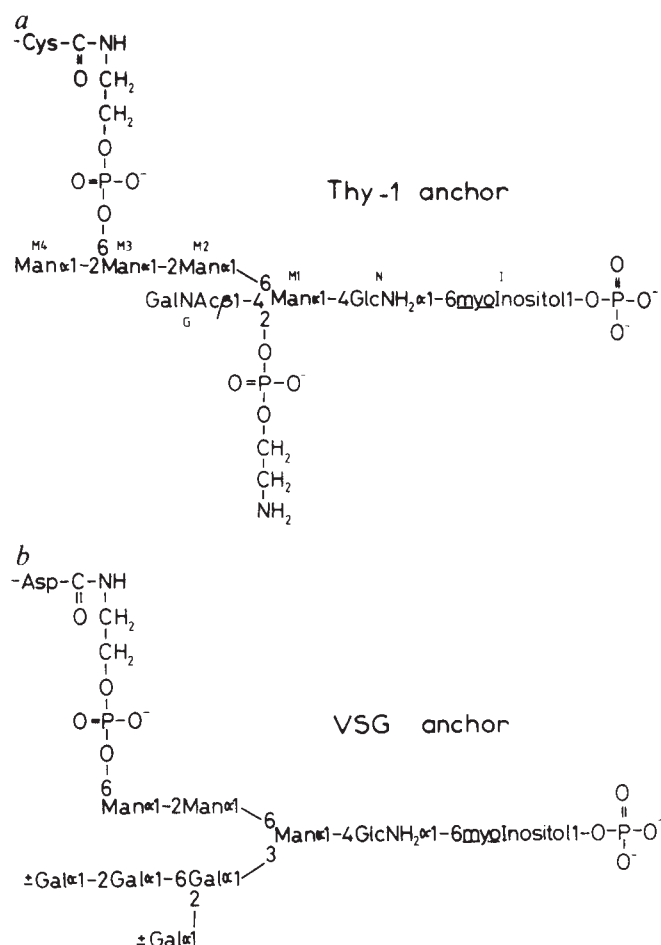


Fig. 2 Composite Structures of the delipidated GPI anchors of rat brain Thy-1 (*a*) and *T. brucei* VSG MITat.1.4 (*b*). *a*, The abbreviations associated with each residue in the Thy-1 structure are those used in the legend to Fig. 1. The proposed Thy-1 structure is homogeneous except for Man-4 which was absent in 29% of the molecules and determined by Bio Gel P4 chromatography (Table 2) and is consistent with NMR, GC-MS and exoglycosidase digestion data. *b*, The VSG GPI structure as reported elsewhere⁷.

compositional analysis since the mannose-6-phosphate bond is stable in the hydrolysis conditions used for carbohydrate analysis). The extra mannose residue (Man-4) that is absent from the VSG anchor is also unlikely to be present in rat thymocyte Thy-1 GPI since this gave a mannose composition of 2 mol mol⁻¹ of anchor². Thus the differences in the structure of the rat brain Thy-1 and trypanosome VSG anchors establish the existence of side chain variation that is specific to species or protein molecular type and the composition data on Thy-1 suggests that tissue-specific side chain variation can also occur. Considerable variation in the nature of acyl or alkyl groups in the phosphatidylinositol moiety has also been established by composition⁶ and other analysis^{8,14}.

The survival of African trypanosomes in the blood is dependent on the integrity of their VSG coats¹⁵ and the GPI anchor has therefore been suggested as a potential target for chemotherapy using specific inhibitors GPI biosynthesis¹⁶. Our results strongly suggest that the backbone structure of the GPI anchor will be common to both parasite and host and this limits the potential usefulness of designing GPI-specific drugs. It may, however, only be necessary to inhibit GPI synthesis transiently to render the VSG coat permeable to lytic factors of host serum. Thus drugs mediating short term inhibition of GPI synthesis might be lethal to the parasite but not the host. An alternative strategy is to design drugs to target the α -galactose

branch of the *T. brucei* GPI anchor. A structural space-filling role in the VSG coat has been suggested for this galactose side chain¹⁷ and our results suggest it could be parasite-specific.

This work was supported in part by Monsanto and the Wellcome Trust.

Received 25 February; accepted 14 April 1988.

1. Ferguson, M. A. J., Low, M. G. & Cross, G. A. M. *J. biol. Chem.* **260**, 14547-14555 (1985).
2. Tse, A. G. D., Barclay, A. N., Watts, A. & Williams, A. F. *Science* **230**, 1003-1008 (1985).
3. Futerman, A. H., Low, M. G., Ackermann, K. E., Sherman, W. L. & Silman, I. *Biochem. biophys. Res. Commun.* **129**, 312-317 (1985).
4. Haas, R., Brandt, P. T., Knight, J. & Rosenberry, T. L. *Biochemistry* **25**, 3098-3105 (1986).
5. Low, M. G. *Biochem. J.* **244**, 1-13 (1987).
6. Ferguson, M. A. J. & Williams, A. F. *Rev. Biochem.* **57**, 285-320 (1988).
7. Ferguson, M. A. J., Homans, S. W., Dwek, R. A. & Rademacher, T. W. *Science* **239**, 753-759 (1988).
8. Roberts, W. L., Kim, B. H. & Rosenberry, T. L. *Proc. natn. Acad. Sci. U.S.A.* **84**, 7817-7821 (1987).
9. Medof, M. E., Walter, E. T., Roberts, W. L., Haas, R. & Rosenberry, T. L. *Biochemistry* **25**, 6740-6747 (1986).
10. Stahl, N., Borchelt, D. R., Hsiao, K. & Prusiner, S. B. *Cell* **51**, 229-240 (1987).
11. Ogata, S., Hayasni, Y., Yakami, N. & Ikehara, Y. *J. biol. Chem.* (in the press).
12. Williams, A. F., Tse, A. G. D. & Gagnon, J. *Immunogenetics* **27**, 174-179 (1988).
13. Fatemi, S. H., Haas, R., Jentoft, N., Rosenberry, T. L. & Tartakoff, A. M. *J. biol. Chem.* **262**, 4728-4732 (1987).
14. Roberts, W., Myher, J. J., Kuksis, A. & Rosenberry, T. L. *Biochem. biophys. Res. Commun.* **150**, 271-277 (1988).
15. Cross, G. A. M. *Phil. Trans. R. Soc. B307*, 3-12 (1984).
16. Turner, M. J. *Br. med. Bull.* **41**, 137-143 (1985).
17. Ferguson, M. A. J., Homans, S. W., Dwek, R. A. & Rademacher, T. W. *Biochem. Soc. Trans.* (in the press).
18. De Bruyn, A., Anteunis, M. & Verhegge, G. *Acta scientia indica* **1**, 83-88 (1975).
19. Ernst, R. R., Bodenhausen, G. & Wokaun, A. *Principles of NMR in One of Two Dimensions* (Clarendon, Oxford, 1987).
20. Campbell, D. G., Gagon, J., Reid, K. B. M. & Williams, A. F. *Biochem. J.* **105**, 15-30 (1981).
21. Low, M. G. *Meih. Enzym.* **71**, 741-746 (1981).

Cingulin, a new peripheral component of tight junctions

Sandra Citti*†, Helena Sabanay†, Ross Jakes*, Benjamin Geiger† & John Kendrick-Jones*

* MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

† Department of Chemical Immunology, The Weizmann Institute of Science, Rehovot 76100, Israel

The tight junction (*zonula occludens*), a belt-like region of contact between cells of polarized epithelia, serves as a selective barrier to small molecules and as a total barrier to large molecules^{1,2}, and is involved in the separation between luminal and basolateral compartments of the epithelium^{3,4}. In the electron microscope, tight junctions show focal regions of apparent fusion between the adjoining cell membranes, and freeze-fractured membranes display an elaborate network of branching and anastomosing strands^{1,5-8}. Very little is known about the molecular composition and architecture of tight junctions. The first specific *zonula occludens*-associated protein, designated ZO-1, has recently been identified in mammalian epithelial and endothelial cells⁹. Here we describe the identification and purification of a new component of this junctional complex in avian brush-border cells, which we name cingulin. Cingulin is an acidic, heat-stable protein, with a highly elongated shape. Immunofluorescence and immunoelectron microscopy of brush-border cells with anti-cingulin antibodies show that cingulin is localized in the apical zone of the terminal web, at the endofacial surfaces of the *zonula occludens*.

During the preparation of antibodies against brush-border myosin, we obtained monoclonal antibodies reactive to a new protein present in the epithelial cells. We initially localized this protein in whole cells and isolated brush borders by immunofluorescence microscopy (Fig. 1). The brush border consists of a tightly packed array of digitiform projections (microvilli) containing bundles of crosslinked actin filaments, and an underlying region, the terminal web, containing actin, α -actinin, vin-

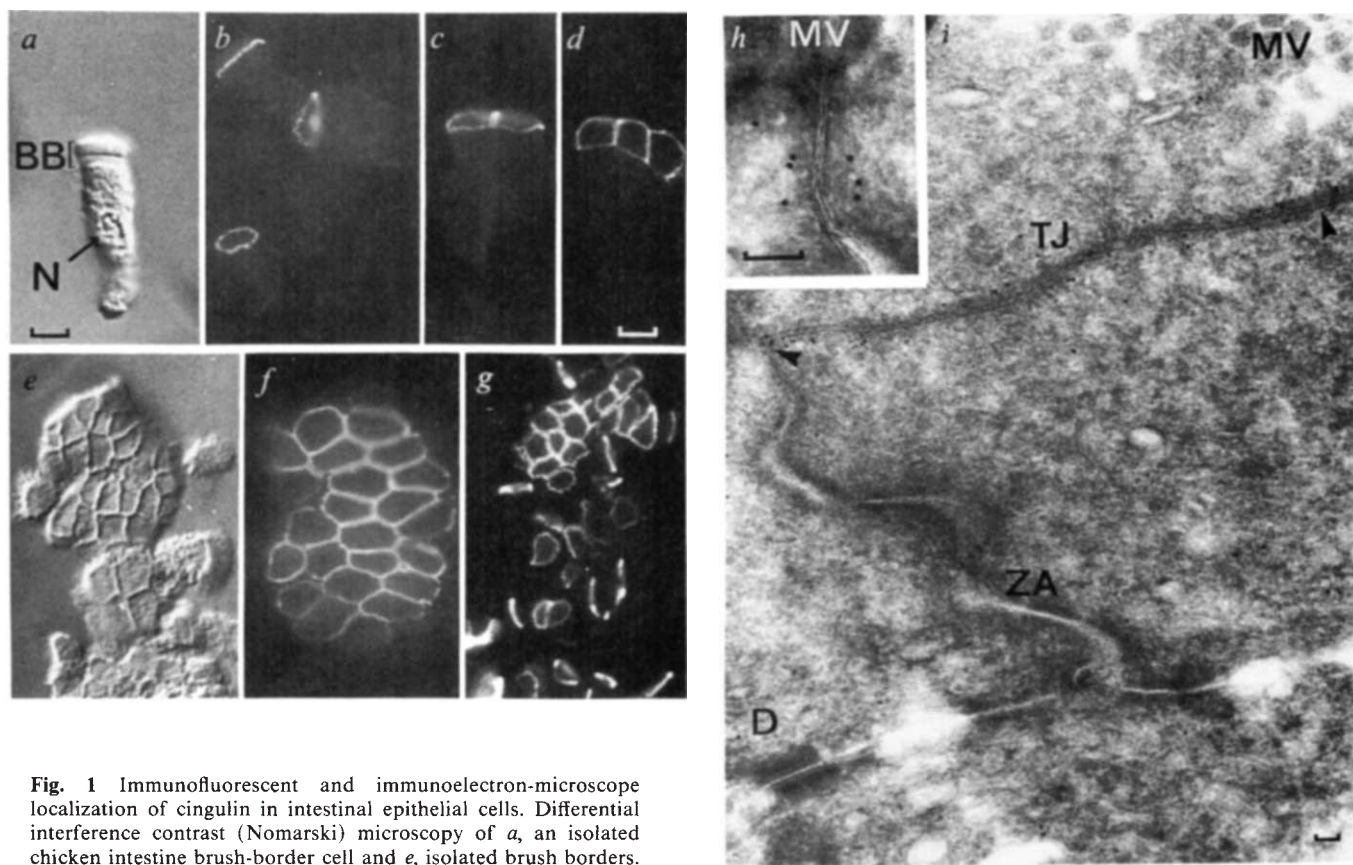


Fig. 1 Immunofluorescent and immunoelectron-microscope localization of cingulin in intestinal epithelial cells. Differential interference contrast (Nomarski) microscopy of *a*, an isolated chicken intestine brush-border cell and *e*, isolated brush borders. Immunofluorescence microscopy using anti-cingulin monoclonal antibodies on *d-f*, brush-border cells and *g*, isolated brush borders. *h, i*, Anti-cingulin immunogold-labelled ultrathin frozen sections of chicken intestine. BB, brush border; N, nucleus; MV, microvilli; TJ, tight junction (arrowheads, tight-junctional area); ZA, zonula adherens; D, desmosome. Scale bars, *a-g*, 10 μm ; *h, i*, 100 nm.

Methods. Monoclonal antibodies directed against a minor contaminant (cingulin) in myosin preparations (estimated <1%), which did not cross react with either brush border myosin rod or myosin subfragment-1 (S1), were detected while screening for anti-myosin S1 antibodies²⁹. Distinct NSO hybridoma lines were denoted Ci6, Ci7, Ci12 and Ci14. Antibodies (IgGs) were purified as described²⁹. Secondary antibodies for immunofluorescence were affinity purified FITC-labelled goat anti-mouse antibodies (Miles). Isolated epithelial cells³⁰ were washed and resuspended in phosphate-buffered saline (PBS), fixed with 3.7% formaldehyde, treated with 1 mg ml⁻¹ collagenase (Sigma), sheared through a 19-gauge needle and washed in PBS. Isolated brush borders were prepared by a modification of the method³¹ (Paul Matsudaira, personal communication). Suspensions were placed onto polylysine coated slides, permeabilized in acetone at -20 °C for 10 min and dried. For immunofluorescent staining, the antibodies Ci6, Ci7, Ci12 or Ci14 were used (ascites fluid, diluted 1:500 to 1:2,000 into PBS containing 0.1% bovine serum albumin BSA). Secondary antibody was used at a dilution of 1:20 in PBS containing 0.1% BSA. Incubations were 1 h at room temperature. All antibodies gave the same immunofluorescent staining pattern. Controls included staining with mouse monoclonal antibodies towards brush-border myosin, mouse preimmune serum and second antibody only. Specimens were examined with a Zeiss WL epifluorescence microscope equipped with a HBO 50 W high-pressure mercury light and selective FITC filter (487710), and using a Zeiss planachromat 100 $\times$ N.A. 1.25 or Neofluar 100 $\times$ N. A. 1.3 objective lenses. Micrographs were taken using Ilford XP1-400 ASA film. For immunolabelling with anti-cingulin, freshly excised chicken intestine was cut into small pieces, fixed in 3% paraformaldehyde (PFA) in 0.2 M cacodylate buffer containing 10 mM CaCl₂, pH 7.3 for 1 h, embedded in 10% gelatine (in the same buffer) and stored at 4 °C in 0.2 M cacodylate buffer containing 0.1% PFA. Sections were obtained and immunolabelled using Protein-A attached at 10-nm gold particles (Janssen Pharmaceutica) as secondary reagent³². Primary antibody was a 1:1:1 cocktail of purified anti-cingulin antibodies (Ci6, Ci7 and Ci14), at a concentration of ~100 $\mu\text{g ml}^{-1}$.

Labelled sections were examined in a Phillips EM410 electron microscope, operated at 80 kV.

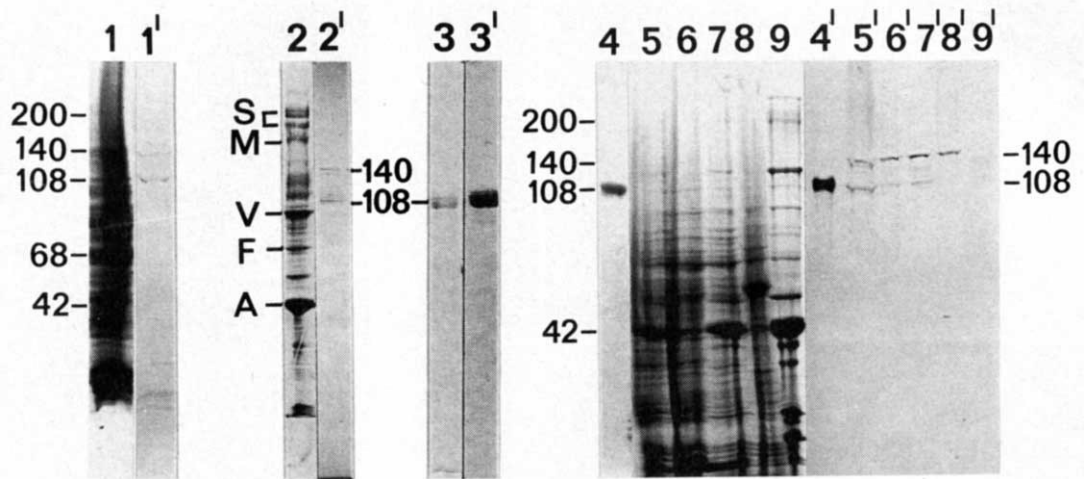
culin, myosin, tropomyosin, the brush-border-specific isoform of spectrin (TW260/240), and intermediate filaments, interwoven in a complex cytoskeletal meshwork (see refs 10,11 for reviews). In the terminal web, the lateral membranes display a characteristic tripartite junctional complex, containing *zonula occludens* (in the apical zone), *zonula adherens* (in the *adhaerens*, or intermediate zone), and spot desmosomes (in the basal zone)^{2,12}. When we labelled isolated brush-border cells with our monoclonal antibodies, we observed a prominent, ring-like immunofluorescent pattern localized in the apico-lateral area of the terminal web. No labelling was detectable in the microvilli, or in other regions of the cell (Fig. 1*b*). We subsequently characterized the antigen recognized by these antibodies biochemically (see below) and, because of its characteristic immunofluorescent

pattern, we named it cingulin (from the latin *cingere*, to encircle)¹³.

When anti-cingulin-labelled intestinal cells still attached to each other in a sheet are viewed *en face*, there is a honeycomb-like immunofluorescent pattern in the regions of contact between neighbouring cells (Fig. 1*c, d, f*). Isolated brush borders are labelled in a similar polygonal pattern, indicating that the antigen recognized by our antibodies is not lost after cell disruption (Fig. 1*g*). In its general appearance, the labelling pattern with anti-cingulin is reminiscent of that obtained with antibodies directed towards *zonula adherens*-specific cytoskeletal components, such as α -actinin and vinculin¹⁴.

To determine which junctional element within the complex associates with cingulin, we prepared ultrathin frozen sections

Fig. 2 Analysis of cingulin by PAGE and immunoblot. Lanes 1', 2', 3', immunoblots with anti-cingulin monoclonal antibodies. Lanes 4', 5', 6', 7', 8', 9', immunoblots with anti-cingulin polyclonal antiserum. Lane 1, whole brush-border cell (crude extract); 2, isolated brush borders; S, brush-border spectrin-like molecules (220–260K); M, myosin heavy chain (200K); V, villin (95K); F, fimbrin (68K); A, actin (42 KD); 3, 4, purified cingulin (≈ 2 and $\approx 1 \mu\text{g}$); 5, chicken intestine; 6, liver; 7, kidney; 8, pancreas; 9, gizzard.



Methods. Freshly prepared whole brush-border cells, isolated brush borders and freshly excised (or frozen ground) tissues from chicken were rapidly transferred into SDS sample buffer (containing protease inhibitors³⁰) and immediately boiled for 2 min. Great care was taken in the preparation of cells and tissues to minimize proteolysis, and a wide spectrum of protease inhibitors were included³⁰. In gels of whole tissues, cells and brush borders, each lane was overloaded to maximize the amount of cingulin or other possible crossreacting species, when the bands were transferred onto nitrocellulose. Chicken intestine brush-border cells prepared as described previously³⁰ were homogenized in 0.6 M NaCl, 0.3 M sucrose, 5 mM EGTA, 25 mM Tris-HCl, pH 7.5, 10 mM sodium phosphate, pH 7.5, 0.1 mM dithiothreitol (DTT), 0.1 mM phenylmethylsulphonyl fluoride (PMSF) and other protease³⁰ at 4°C. Extract supernatant (after centrifugation at 100,000g for 1 h) was fractionated with ammonium sulphate, and the 0–40% wash pellet dissolved and dialysed against 0.3 M NaCl, 5 mM MgCl₂, 0.5 mM EGTA, 20 mM Tris-HCl, pH 7.5, 0.1 mM DTT, and heated at $\approx 100^\circ\text{C}$ for ≈ 5 min. Denatured proteins were removed by centrifugation 28,000g 15 min). The supernatant (containing all the immunoreactive cingulin) was dialysed into 20 mM Tris-HCl, pH 7.5, 1 mM EGTA, 0.1 mM DTT, and chromatographed on a Pharmacia Mono-Q HPLC anion-exchange column, equilibrated in the same solution, and cingulin eluted as a symmetrical peak at ≈ 500 mM NaCl. Polyclonal anti-cingulin antiserum was raised in rabbit by subcutaneous injection of $\approx 50 \mu\text{g}$ purified cingulin, at 2-month intervals. Polyacrylamide gradient mini slab-gels (5–20%) were run as described³³, transferred onto nitrocellulose and immunostained as described³⁴. Efficiency of transfer onto nitrocellulose was checked by staining blots with 0.1% amido black. For the immunoblots, primary antibody was either a 1:1:1 mixture of monoclonal antibodies Ci6, Ci12 and Ci14 (ascites fluid), or polyclonal rabbit antiserum, diluted 1:100 in PBS containing 2% BSA. Secondary antibody was horseradish peroxidase-labelled, rabbit anti-mouse immunoglobulin, or horseradish peroxidase-labelled or goat anti-rabbit immunoglobulin antibody (Dako-Patts), diluted 1:400 in PBS/BSA²⁹.

of chicken intestine and subjected them to indirect immunogold labelling with our anti-cingulin antibodies. We consistently observed labelling near the cytoplasmic faces of the membrane (mean distance of the gold particles from the membrane midline ~ 40 nm) in the apical zone of the terminal web, which is in the region of the tight junction (Fig. 1*h*, *i*). Examination of the membrane configurations along the labelled areas at high magnification revealed loop-like structures, with an apparent fusion of the outer leaflets of the plasma membranes, typical of tight junctions (Fig. 1*h*). The labelling with anti-cingulin was abundant not only in the tight junctions between neighbouring columnar epithelial cells but also in junctions formed with goblet cells (not shown). We did not detect labelling at the level of the *zonulae adhaerentes* or the desmosomes (Fig. 1*i*). Cingulin has been detected in various epithelial and endothelial cells, and its distribution and further studies on its subcellular localization will be described elsewhere (S.C., H.S., J.K.-J. and B.G., manuscript in preparation).

Western blot analysis of whole epithelial cell homogenates following SDS-PAGE reveals that the anti-cingulin antibodies crossreact with two polypeptides of apparent relative molecular mass (M_r) $\sim 140,000$ (140K) (cingulin-140) and ~ 108 K (cingulin-108) (Fig. 2, lanes 1, 1'). A similar pattern is obtained on western blots of isolated brush borders (Fig. 2, lane 2, 2'). Because of the extremely low amounts of cingulin present in whole tissues, it was difficult to detect the two bands in blots of whole intestine using monoclonal antibodies. We therefore used a rabbit polyclonal antiserum raised against the purified protein to obtain greater sensitivity and to characterize the cingulin-immunoreactive polypeptides in various tissues, including intestine.

As shown in Fig. 2 (lanes 5–9), polypeptides of similar M_r (140K and 108K) to those detected in intestine are labelled by

the polyclonal antiserum in liver, kidney and pancreas (and also in other tissues characterized by polarized epithelium; data not shown). Blots of gizzard smooth muscle tissue, which does not contain tight junctions (except for those present in vascular endothelial cells), are not labelled by the anti-cingulin antibodies (Fig. 2, lane 9). In summary, neither the monoclonal nor the polyclonal antibodies react with major polypeptides of M_r other than 140K and 108K in whole tissues, isolated cells and brush borders (Fig. 2).

Both cingulin-140 and cingulin-108 could be extracted from brush-border cells at approximately physiological ionic strength, and we purified cingulin-108 to $>95\%$ homogeneity (Fig. 2, lanes 3 and 4) by a procedure involving heat treatment of brush-border extracts (cingulin is heat-stable) and anion-exchange chromatography. We detected cingulin-140 polypeptide only in crude tissue samples, and extract and therefore we could not purify it. The relationship between the cingulin-140 and the cingulin-108 polypeptides, which are recognized by the same antibodies, is not yet clear. Further studies are necessary to establish whether the presence of these two forms of cingulin, with different electrophoretic mobility, result from the presence of multiple protein isoforms, from proteolysis, or from other types of chemical modification, such as sulphhydryl oxidation/reduction.

We determined the amino-acid composition of cingulin-108 (Table 1). We found no obvious relationship between cingulin and several microfilament-associated proteins, such as caldesmon¹⁵ ($M_r \sim 135$ K), myosin rod¹⁶ (~ 120 K), vinculin^{17,18}, (130K), α -actinin¹⁹ (95K) and spectrin²⁰ (240K subunit). Moreover, immunochemical studies (Western blotting and immunofluorescence) indicate that cingulin is antigenically distinguishable from myosin and its fragments, spectrins (see also Fig. 2, lanes 2, 2'), skeletal muscle α -actinin, and smooth muscle

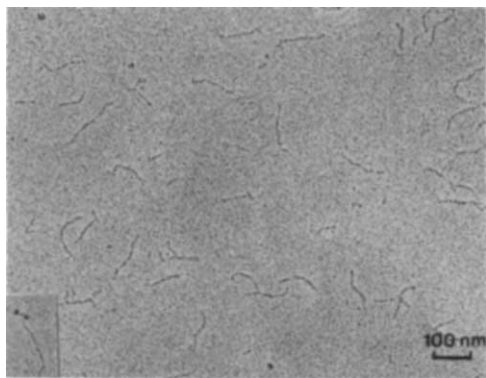


Fig. 3 Molecular shape of purified cingulin-108. Electron micrograph showing a field of purified cingulin molecules, after rotary shadowing with platinum. Insert, one shadowed brush-border myosin molecule, at the same magnification as cingulin, for comparison. Scale bar, 100 nm.

Methods. For rotary shadowing, 150 μ l purified protein (~ 0.6 mg ml $^{-1}$ in 500 mM NaCl, 25 mM Tris-HCl, pH 7.5, 1 mM EGTA, 0.1 mM DTT) was diluted with an equal volume of glycerol (final 50% glycerol, vol/vol), mixed and immediately sprayed onto freshly cleaved mica, using a manually operated glass atomizer. The mica was rotary shadowed with platinum at an angle of 5° and at a pressure of $\sim 1.10^{-5}$ atm (ref. 35), and coated with carbon. Replicas were floated off onto distilled water and picked up with bare 400-mesh grids. Grids were examined in a Phillips EM 400 electron microscope, operated at 80 kV.

caldesmon, vinculin and myosin light-chain kinase^{14,15,22-24}. Although the apparent high ratio of charged to apolar residues suggests a protein with a high α -helical content²¹, the presence of proline residues (18 mol mol $^{-1}$) excludes the possibility that cingulin is a fully α -helical protein (Table 1). Note that cingulin contains a number of cysteine residues, and our preliminary studies indicate that the molecule is extremely susceptible to sulphhydryl crosslinking.

Electron microscopy of rotary-shadowed, purified cingulin (108K polypeptide) reveals elongated, rod-like molecules, with a mean contour length of 130 nm (s.d., 32 nm; n , 381), and a width of 2–3 nm, very similar to that of brush-border myosin rods (Fig. 3). The length of cingulin molecules is in good agreement with that (122 nm) predicted for a α -helical, coiled coil protein of chain weight 108K (ref. 25). The cingulin rods are often curved or kinked at various angles (ranging between $\sim 25^\circ$ and $\sim 150^\circ$) and at variable positions along their length, suggesting some degree of flexibility along the molecule. Interestingly, our results show that cingulin shares some physical-chemical and general structural features (for example, elongated shape, heat stability) with characterized microfilament-associated proteins, such as myosin rod, tropomyosin, the spectrins/fodrin and caldesmon^{15,26-28}. By gel filtration using these elongated proteins as standards, cingulin elutes in a position corresponding to a dimer, composed of two units of 108K and contour length ~ 130 nm. Further structural and functional homologies between cingulin and these proteins remain to be established.

Attempts to understand structure-function relationships in tight junctions at a molecular level have been seriously hampered by the limited amount of data available concerning their fine molecular architecture and the identity of their various constituents. Recently, the immunochemical identification of the first *zonula occludens*-specific protein, denoted ZO-1 (225K), was reported⁹, using a tight-junction-rich fraction from liver tissue as an immunogen to raise specific antibodies. The anti-ZO-1 antibodies stain epithelial and endothelial cells at the light- and electron-microscope level, giving patterns similar to those observed here with anti-cingulin antibodies. These data suggest

Table 1 Amino-acid composition of chicken brush-border cingulin

Amino acid	Mol mol $^{-1}$	Amino acid	Mol mol $^{-1}$
Asp	68	Ile	18
Thr	31	Leu	109
Ser	58	Tyr	6
Glu	260	Phe	8
Pro	18	Lys	75
Gly	43	His	14
Ala	95	Arg	110
Val	34	Cys	8
Met	17		

Protein samples (100–200 pmol) were analysed on a Durrum 500 amino-acid analyser after hydrolysis in 6 N HCl, 0.1% phenol, *in vacuo* at 105 °C. Duplicate samples were taken after 24, 48 and 72 h. Compositions are expressed as moles per mole (assuming that cingulin is a single 110K polypeptide). Cysteine content was determined by performic acid oxidation, using oxidized pig plasma gelsolin (with known cysteine content) as a standard. The composition represents the mean values from four different determinations.

that the two proteins are associated with grossly similar structures at the level of the tight junction. However, the limited information available on the biochemical properties of ZO-1 makes it difficult to compare further it with cingulin.

Our present results strongly suggest that cingulin is peripherally associated with the tight junction membrane, and is not an integral membrane component. This is based on the observation that cingulin can be extracted from brush borders under conditions of near physiological ionic strength, without detergents, and on the immunoelectron-microscope localization, showing that cingulin is close to but not co-localized with the junctional membrane itself. The nature of the tight-junction elements to which cingulin specifically binds and its involvement in maintaining the structure of tight junctions are yet to be resolved.

We thank Talila Volk and Tova Volberg (Weizmann Institute) and Alan Weeds (LMB) for helpful advice and discussion, John Kilmartin (LMB) for the FITC-labelled antibody, and the Boehringer Ingelheim Fonds for travel funds. S.C. was supported by a Thomas C. Usher Foundation Fellowship and by the European Molecular Biology Organization. B.G. was supported by the Muscular Dystrophy Association and the Rockefeller-Weizmann Foundation. B.G. is an Erwin Neter Professor of Cell and Tumor Biology.

Received 8 March; accepted 21 April 1988.

- Farquhar, M. G. & Palade, G. *J. cell Biol.* **17**, 375–412 (1963).
- Staehelein, L. A. *Int. Rev. Cytol.* **39**, 191–283 (1974).
- Simons, K. & Fuller, S. D. A. *Rev. Cell Biol.* **1**, 243–288 (1985).
- Van Meer, G. & Simons, K. *EMBO J.* **5**, 1455–1464 (1986).
- Goodenough, D. A. & Revel, J. P. *J. cell Biol.* **45**, 272–290 (1970).
- Kachar, B. & Reese, T. S. *Nature* **296**, 464–466 (1982).
- Pinto da Silva, P. & Kachar, B. *Cell* **28**, 441–450 (1982).
- Stevenson, B. R. & Goodenough, D. A. *J. cell Biol.* **98**, 1209–1221 (1984).
- Stevenson, B. R., Siliciano, J. D., Mooseker, M. S. & Goodenough, D. A. *J. cell Biol.* **103**, 755–766 (1986).
- Matsudaira, P. T. & Burgess, D. R. *Cold Spring Harb. Symp. quant. Biol.* **46**, 845–854 (1982).
- Mooseker, M. S. A. *Rev. Cell Biol.* **1**, 209–241 (1985).
- Hull, B. E. & Staehelein, L. A. *J. cell Biol.* **81**, 67–82 (1979).
- Citi, S. & Kendrick-Jones, J. *J. cell Biol.* **103**, Pt2, 394a (1986).
- Geiger, B., Dutton, A. H., Tokuyasu, K. T. & Singer, S. J. *J. cell Biol.* **91**, 614–628 (1981).
- Marston, S. B. & Smith, W. J. *J. Muscle Res. Cell Motil.* **6**, 669–708 (1985).
- Lowey, S., Slayter, H. S., Weeds, A. & Baker, H. *J. molec. Biol.* **42**, 1–29 (1969).
- Jockusch, B. M. & Isenberg, G. *Cold Spring Harb. Symp. quant. Biol.* **46**, 613–623 (1981).
- Geiger, B. in *International Cell Biology 1980–1981* 761–773 (Springer, Berlin 1981).
- Ebashi, S. & Ebashi, F. *J. biochem. Tokyo* **58**, 7–12 (1965).
- Glenney, J. R., Glenney, P. & Weber, K. *J. biol. Chem.* **257**, 9781–9787 (1982).
- Cohen, C. & Parry, D. A. D. *Trends biochem. Sci.* **11**, 245–248 (1986).
- De Lanerolle, P., Adelstein, R. S., Feramisco, J. R. & Burridge, K. *Proc. natn. Acad. Sci. U.S.A.* **78**, 4738–4742 (1981).
- Glenney, J. R., Glenney, P. & Weber, K. *J. cell Biol.* **96**, 1491–1496 (1983).
- Hirokawa, N., Cheney, R. E. & Willard, M. *Cell* **32**, 953–965 (1983).
- Stewart, M. & Edwards, P. *FEBS Lett.* **168**, 75–78 (1984).
- Caspar, D. L. D., Cohen, C. & Longley, W. *J. molec. Biol.* **41**, 87–107 (1969).
- Glenney, J. R., Glenney, P. & Weber, K. *J. molec. Biol.* **167**, 275–293 (1983).

28. Bretscher, A. *J. biol. Chem.* **259**, 12873-12880 (1984).
29. Citi, S. & Kendrick-Jones, J. *Eur. J. Biochem.* **165**, 315-325 (1987).
30. Citi, S. & Kendrick-Jones, J. *J. molec. Biol.* **188**, 369-382 (1986).
31. Bretscher, A. & Weber, K. *J. cell Biol.* **86**, 335-340 (1980).
32. Volk, T. & Geiger, B. *J. cell Biol.* **103**, 1441-1450 (1986).
33. Matsudaira, P. T. & Burgess, D. R. *Analyt. Biochem.* **87**, 386-396 (1978).
34. Burnette, W. N. *Analyt. Biochem.* **112**, 195-203 (1981).
35. Burke, B. E. & Shotton, D. M. *EMBO J.* **1**, 505-508 (1982).

Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings

Norbert Schnell*, Karl-Dieter Entian*,
Ursula Schneider†, Friedrich Götz†, Hans Zähner†,
Roland Kellner‡ & Günther Jung‡§

* Medizinisch-Naturwissenschaftliches Forschungszentrum,
University of Tübingen, D-7400 Tübingen, FRG

† Institut für Biologie II, University of Tübingen,
D-7400 Tübingen, FRG

‡ Institut für Organische Chemie, University of Tübingen, D-7400
Tübingen, FRG

The genetic basis for the biosynthesis of large polypeptide antibiotics such as nisin has not been explained so far. We show here that the structural gene *epiA* encoding the antibiotic epidermin^{1,2} from *Staphylococcus epidermidis* is located on a 54-kilobase plasmid and codes for a 52-amino-acid prepeptide, which is processed to the tetracyclic 21-peptide amide antibiotic. The mature sequence of epidermin corresponds to the C-terminal 22-peptide segment of pre-epidermin and contains the precursor amino acids Ser, Thr and Cys, from which the unusual amino-acid constituents are derived. The more lipophilic epidermin is cleaved at a hydrophilic turn between Arg⁻¹ and Ile⁺¹ from the N-terminal segment -30 to -1, which probably assumes a partially amphiphilic α -helix conformation. We propose that the N-terminus (-30 to -1) plays a cooperative role during modification reactions and prevents toxicity of the mature epidermin to the producing strain before the antibiotic is cleaved off and secreted.

Epidermin is a 21-amino-acid peptide amide antibiotic with high antimicrobial activities against pathogenic Gram-positive bacteria such as *Propionibacterium acnes*, staphylococci and streptococci, and hence is of therapeutic value in topical treatment of acne. The complex polypeptide contains four sulphide

rings, consisting of two *meso*-lanthionine (Lan), one (2*S*, 3*S*, 6*R*)-3-methylanthionine (MeLan), and the newly described and unusual amino acid S-(2-aminovinyl)-D-cysteine (D-Cys(Avi)). An additional α,β -unsaturated amino acid, dehydrobutyryne (Dhb) is situated at the tryptic cleavage site of epidermin^{1,2}. Epidermin has some structural similarities to the well-known antibiotic nisin³ that is used in food preservation. In addition to epidermin and nisin, lanthionine-containing polypeptide antibiotics include subtilin⁴, cinnamycin and duramycin⁵, anconin⁶, Ro09-0198 (ref. 7) and Pep5 (ref. 8). We suggest the common name lantibiotics for these polypeptides, because of their particular and unique mode of biosynthesis.

There are two alternative routes of biosynthesis discussed for lantibiotics, namely by multienzyme complexes or by ribosomal biosynthesis. To distinguish between these, we have synthesized a wobbled DNA probe 5'-GTG(A)CAT(G/A)ATG(A)-AAT(C)TT-3', deduced from a suitable pentapeptide segment of the proposed pre-sequence of epidermin: LysPheIleCysThr. This DNA probe was hybridized against chromosomal and plasmid DNA (refs 9-11) from *S. epidermidis* (Tü3298) that harbours a 54-kilobase (kb) plasmid, pEpi32. Restriction analysis of the isolated plasmid gave seven DNA fragments with *EcoRI* (16, 11, 10, 6.5, 5.5, 3.5 and 2.5 kb), nine DNA fragments with *HindIII* (17, 14, 10, 5.3, 2.8, 1.8, 0.8, 0.6 and 0.5 kb) and five DNA fragments with *BamHI* (20, 19, 10, 3 and 1 kb). A hybridizing 5.4-kb *HindIII* fragment was subcloned and rehybridization located the epidermin structure gene *epiA*, within a 2.2-kb *EcoRI/BglII* fragment. As we had a mixture of 24 different 14-mers as a hybridization probe, we applied it in a 30-fold excess as a sequencing primer in accordance with established techniques. DNA sequencing provided a short identifiable region which was used for subsequent priming.

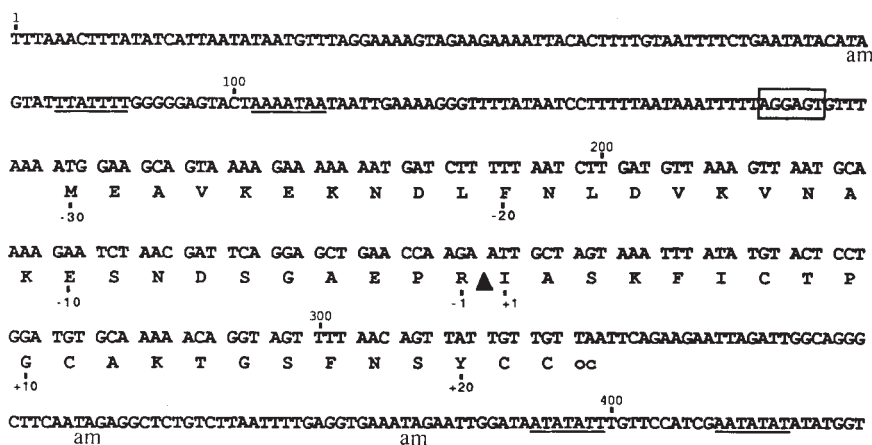
The peptide sequence of epidermin identified the open-reading frame. A single methionine codon is at an appropriate distance to a Shine-Dalgarno sequence. The structural gene of pre-epidermin terminates at the TAA stop codon, hence pre-epidermin consists of 52 amino acids (Fig. 1). It is processed to the epidermin between Arg⁻¹ and Ile⁺¹. From these results we can conclude that pre-epidermin is not a degradation product of any larger protein and show here for the first time that the precursor proteins of the lantibiotics are encoded by a distinct structural gene.

A combination of prediction profiles for secondary structure (α , β , turn), flexibility, hydrophathy, hydrophilicity and helix wheel plot were made using the program HYCON (Fig. 2). A high α -helix probability is predicted for pre-epidermin -30 to -8, whereas the C-terminal region 1-22, comprising pro-epidermin, has very high turn probability. Furthermore, the prediction plot shows clearly that the N-terminus -30 to -1 is highly hydrophilic, whereas the C-terminal region is more

§ To whom correspondence should be addressed.

Fig. 1 Nucleotide sequence of the epidermin structural gene (*epiA*) and deduced amino-acid sequence of pre-epidermin. The Shine-Dalgarno sequence 7 base pairs in front of the ATG codon is boxed, and the proteolytic cleavage site at which the prepeptide is processed is indicated by an arrow. Inverted repeats are underlined and potential stop codons are given (am, amber; oc, ochre).

Methods. Both strands of *epiA* were sequenced using the dideoxy chain termination method¹⁴. The 2.2-kb fragment containing *epiA* was ligated into pUC19 and plasmid sequencing was done as described previously¹⁵. Plasmids were purified by CsCl centrifugation in a mini ultracentrifuge (Beckman T100, rotor TLA 100.2, 80,000 r.p.m., 12 h). The mixed oligonucleotide used for *epiA* cloning was successful as a first primer and additional primers were synthesized on a 380B DNA synthesizer.



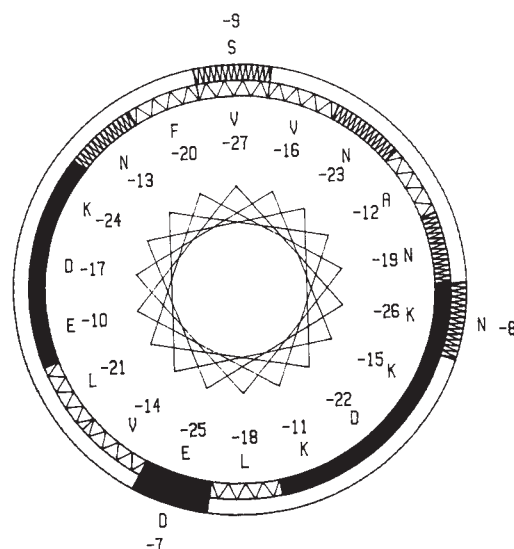
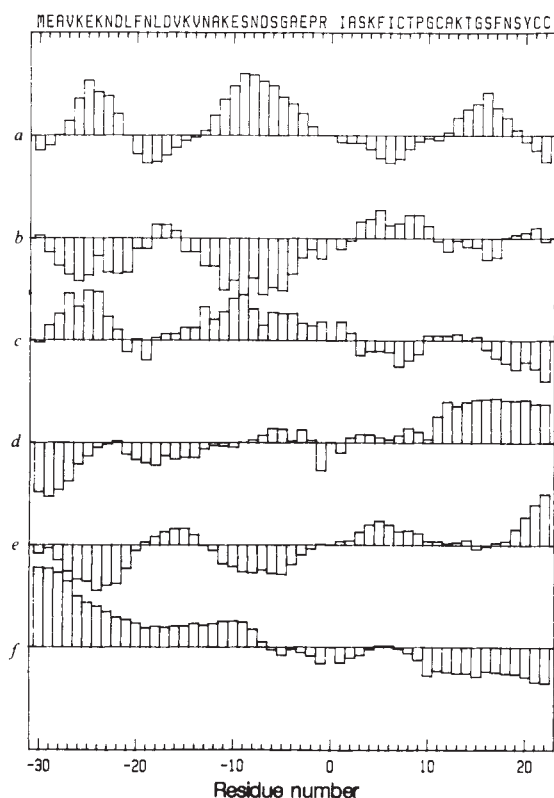


Fig. 2 Prediction plots for pre-epidermin (left). *a*, flexibility¹⁶; *b*, hydropathy¹⁷; *c*, hydrophilicity¹⁸; *d*, propensities for turn; *e*, β -sheet and *f*, α -helix conformation¹⁹. As shown by the helix-wheel plot (right), the N-terminus may partially adopt an amphiphilic α -helical conformation in an appropriate environment, possibly by interaction with the C-terminal pre-epidermin part which becomes highly lipophilic during sulphide ring formation. Modification steps occur at Ser and Thr residues in exposed turn structures.

lipophilic. The N-terminal part -30 to -8 seems to fold partially into an amphiphilic α -helix, as suggested from helix wheel plots.

The N-terminal segment of pre-epidermin -30 to -1 does not contain any cysteine residues, whereas the C-terminal segment 1-22 contains the four cysteine residues involved in sulphide bridge formation. Clearly, sequence -30 to -1 possesses many cleavage sites for endoproteases, whereas even in the pre-epidermin state, sequence 1-22 is resistant to proteolytic degradation. The mature antibiotic epidermin^{1,2} can only be attacked by trypsin at Lys in position 13. The processing site Arg⁻¹-Ile⁺¹ of epidermin is hydrophilic and accessible, owing to turn-forming Pro⁻² residue.

The question arises how the modification and processing steps follow each other, as various enzymatic reactions are necessary. These reactions include (1) modification of the region 1-22, (2) cleavage of the N-terminal prepeptide fragment -30 to -1 and (3) secretion of the mature antibiotic. Processing could occur before or after maturation. Our present hypothesis (Fig. 3) favours the occurrence of enzymatic modifications at the prepeptide before cleavage. Two observations support this view: first, in contrast to the C-terminus, the N-terminus of the pre-epidermin contains mostly hydrophilic amino acid residues, hence increasing propeptide solubility. Secondly, the mature antibiotic is also toxic for the producing strain, and we assume that the partially amphiphilic α -helical N-terminus may inhibit this toxic action until the antibiotic is processed and secreted. An enzyme complex, possibly membrane-associated, could catalyse the elimination of water from Ser and Thr residues in positions 5, 16, 19 and 8, 14, respectively, to form dehydroalanine and dehydrobutyryne residues. Addition of thiol groups of Cys residues in positions 7, 11, 21, 22 to the C=C double bonds, could occur simultaneously or subsequently, yielding the *meso*-lanthionine or (2*S*, 3*S*, 6*R*)-3-methyl-lanthionine bridges. In addition, decarboxylation of residue 22 and double bond formation yields the C-terminal S-(2-aminovinyl)-D-cysteine. The reaction of C-terminally situated cysteine thiol groups with N-terminally located dehydroamino acids occurs with complete stereospecificity, at least in epidermin^{1,2}, nisin and subtilin^{3,4}.

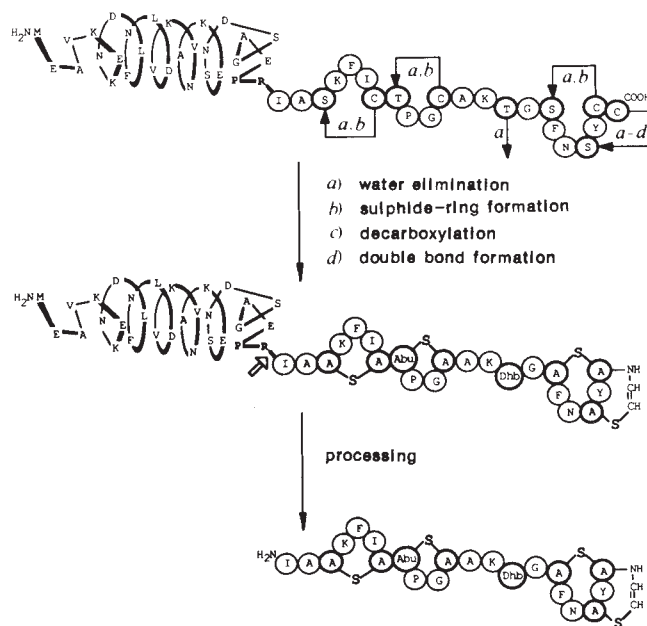


Fig. 3 Hypothetical maturation procedure of epidermin. The translated polypeptide (pre-epidermin) consists of 52-amino-acid residues. Structure predictions indicate a partially α -helical N-terminus from which residues -30 to -10 may form an amphiphilic α -helix conformation. A Pro⁻² turn followed by an arginine residue corresponds to the processing site. (a), A water elimination reaction at Ser and Thr residues. With the exception of Thr⁺¹⁴, water elimination is followed by (b), sulphide-ring formation, (c) describes the decarboxylation reaction, and (d) the double bond formation at the C-terminus. After these modifications, the pre-epidermin is proposed as an intermediate formation, from which the lantibiotic is processed (Abu, aminobutyric acid; Dhb, dehydrobutyryne). The name lantibiotic is suggested for all lanthionine- and/or 3-methylanthionine-containing polycyclic peptide antibiotics, which are synthesized ribosomally through prepeptides.

Therefore during modification these elimination-addition reactions imply a reversal of configuration of the α -carbon atoms at pre-epidermin residues L-Ser (positions 15, 16, 19) and L-Thr (position 8) to give the D-configuration. On the other hand, the L-configuration of the cysteine halves is still maintained.

The four sulphide rings are probably formed subsequently at the same catalytic site, perhaps supported by interaction with the N-terminal amphiphilic α -helix. Chemically and sterically, all first-step dehydration reactions are very unlikely to be catalysed simultaneously as this would cause conformational constrain and yield an unstable polypeptide with five dehydroamino acid residues. Only Thr⁺¹⁴ dehydrates without finding a cysteine. This position (Lys⁺¹³-Dhb⁺¹⁴) constitutes the enzymatic cleavage site^{1,2} at which trypsin inactivates the antibiotic epidermin. During sulphide ring formation, C-terminal rigidity and hydrophobicity increases and may favour interaction of pre-epidermin with the lipid bilayer and induce its translocation. Finally, the hydrophilic α -helical N-terminus -30 to -1 must be cleaved by a specific protease at the characteristic cleavage site described above.

Our results provide more than an experimental basis for elucidating details of lantibiotic biosynthesis; using these modifying enzyme systems, it may become possible to construct proteins and hormones that contain dehydroamino acids and/or thioether bridges replacing disulphides. Such carba-analogues of cystine-bridged hormones are of great interest for studies of structure-activity relationships. Hitherto, these studies were restricted to chemically synthesized, biologically active analogues of small peptide hormones, such as oxytocin and vasopressin¹². Because chemical syntheses are tedious and complicated, particularly for peptide analogues of the size of the

lantibiotic nisin¹³, a biosynthetic approach may become an attractive alternative.

This work was supported by Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie. N.S. was supported by a Dechema grant and K.D.E. is a recipient of a Heisenberg grant of the Deutsche Forschungsgemeinschaft. We thank Professor D. Mecke for discussions, and W. Tröger, F. Gombert and C. C. Jung for programming the HYCON prediction.

Received 7 March; accepted 7 April 1988.

1. Allgaier, H., Jung, G., Werner, R. G., Schneider, U. & Zähler, H. *Angew. Chem.* **24**, 1051-1053 (1985).
2. Allgaier, H., Jung, G., Werner, R. G., Schneider, U. & Zähler, H. *Eur. J. Biochem.* **160**, 9-22 (1986).
3. Gross, E. & Morell, J. L. *J. Am. chem. Soc.* **93**, 4634-4635 (1971).
4. Gross, E., Klitz, H. & Craig, L. C. *Hoppe-Seyler's Z. physiol. Chem.* **354**, 799-801 (1973).
5. Gross, E. *Adv. exp. Med. Biol.* **86b**, 131-153 (1977).
6. Wakamiya, T., Ueki, Y., Shiba, T., Kido, Y. & Motoki, Y. *Tetrahedron Lett.* **26**, 665-668 (1985).
7. Kessler, H., Steuernagel, S., Gillissen, D. & Kamiyama, T. *Helv. chim. Acta* **70**, 726-741 (1987).
8. Sahl, H.-G. *J. Bact.* **162**, 833-836 (1985).
9. Novick, R. P. & Bouanchaud, D. *Ann. N.Y. Acad. Sci.* **182**, 279-294 (1971).
10. Southern, E. J. *molec. Biol.* **98**, 503-517 (1975).
11. Heinrich, P., Rosenstein, R., Böhrer, M., Sonner, P. & Götz, F. *Molec. gen. Genet.* **209**, 563-569 (1987).
12. Hruby, V. J. & Smith, C. W. in *The Peptides, Analysis, Synthesis, Biology* Vol. 8 (eds. Udenfriend, S. & Meienhofer, J.) 77-197 (Academic, Orlando, Florida, 1987).
13. Fukase, K. *et al.* in *Peptide Chemistry 1987* (eds Shiba, T. & Sakakibara, S.) 337-342 (Protein Research Foundation, Osaka, 1988).
14. Sanger, F., Coulson, A. R., Barell, B. G., Smith, A. J. & Roe, B. A. *J. molec. Biol.* **143**, 161-178 (1980).
15. Hattori, T. A. & Sakaki, Y. *Analyt. Biochem.* **152**, 232-238 (1984).
16. Karplus, P. A. & Schulz, G. E. *Naturwissenschaften* **72**, 212-213 (1985).
17. Kyte, J. & Doolittle, R. F. *J. molec. Biol.* **157**, 105-132 (1982).
18. Hopp, T. P. & Woods, K. R. *Molec. Immun.* **20**, 483-489 (1983).
19. Garnier, J., Osguthorpe, D. J. & Robson, B. *J. molec. Biol.* **120**, 97-120 (1978).

Human immunodeficiency virus induces phosphorylation of its cell surface receptor

Alan P. Fields, Daniel P. Bednarik, Allan Hess & W. Stratford May

The Johns Hopkins Oncology Center, 424 North Bond Street, Baltimore, Maryland 21231, USA

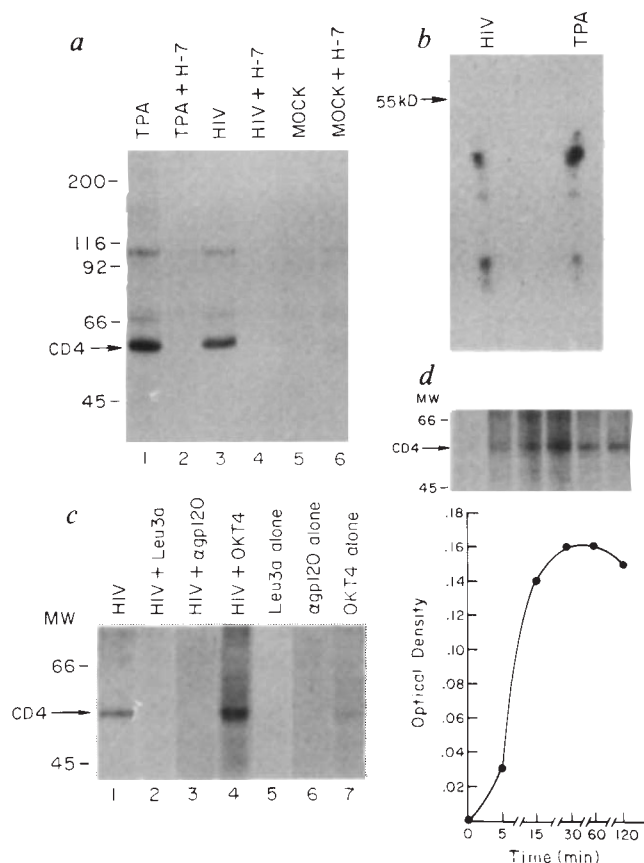
AIDS is an immunoregulatory disorder characterized by depletion of the CD4⁺ helper/inducer lymphocyte population¹. The causative agent of this disease is the human immunodeficiency virus, HIV²⁻⁴, which infects CD4⁺ cells and leads to cytopathic effects characterized by syncytia formation and cell death^{5,6}. Recent studies have demonstrated that binding of HIV to its cellular receptor CD4 is necessary for viral entry^{7,8}. We find that binding of HIV to CD4 induces rapid and sustained phosphorylation of CD4 which could involve protein kinase C. HIV-induced CD4 phosphorylation can be blocked by antibody against CD4 and monoclonal antibody against the HIV envelope glycoprotein gp120, indicating that a specific interaction between CD4 and gp120 is required for phosphorylation. Electron microscopy shows that a protein kinase C inhibitor does not impair binding of HIV to CD4⁺ cells, but causes an apparent accumulation of virus particles at the cell surface, at the same time inhibiting viral infectivity. These results indicate a possible role for HIV-induced CD4 phosphorylation in viral entry and identify a potential target for anti-viral therapy.

Cell surface expression of CD4 can be modulated by both HIV^{5,6} and 12-*o*-tetradecanoyl-phorbol-13-acetate (TPA)⁹⁻¹¹. TPA-induced down-modulation is accompanied by CD4 phosphorylation¹⁰. Therefore we have tested whether HIV could induce CD4 phosphorylation. Cloned human CD4⁺ lym-

phocytes were equilibrated with [³²P]orthophosphate, incubated with either HIV or TPA, and the phosphorylation status of CD4 monitored by immunoprecipitation (Fig. 1a). When cells were treated for ten minutes with 100 ng ml⁻¹ TPA, CD4 (relative molecular mass *M_r* = 55,000) was phosphorylated (Fig. 1a, lane 1). Exposure of cells to HIV for 30 min also resulted in CD4 phosphorylation (Fig. 1a, lane 3). No phosphorylation was observed using a mock inoculum devoid of virus (Fig. 1a, lane 5), demonstrating that phosphorylation is dependent upon the presence of HIV in the inoculum.

TPA-induced CD4 phosphorylation is thought to be mediated by activated protein kinase C. To test whether this enzyme could be involved in HIV-induced CD4 phosphorylation, cells were treated with the protein kinase C inhibitor (1-5-isoquinolinesulphonyl-2-methylpiperazine dihydrochloride; H-7) before incubation with either TPA or HIV. At this concentration, H-7 shows no cellular toxicity and selectively inhibits protein kinase C¹². Both TPA- and HIV-induced CD4 phosphorylation was abolished by H-7 (Fig. 1a, lanes 2 and 4 respectively). To determine whether the same site(s) on CD4 were phosphorylated in response to HIV and TPA, comparative peptide mapping was carried out on ³²P-labelled CD4 (Fig. 1b). The phosphopeptide patterns were identical, irrespective of whether phosphorylation was induced with HIV or TPA, indicating that identical site(s) on CD4 are phosphorylated in both cases. These findings suggest that activated protein kinase C may be involved in both the HIV and TPA responses.

To determine whether HIV-mediated CD4 phosphorylation resulted from the direct binding of HIV to cell-surface CD4, we examined the effect of antibodies directed against either CD4 (Leu3a and OKT4) or the HIV gp120 envelope protein (anti-gp120) on HIV-induced CD4 phosphorylation. When cells were incubated with Leu3a and then exposed to HIV (Fig. 1c, lane 2), or when HIV was incubated with anti-gp120 before incubation with cells (Fig. 1c, lane 3), CD4 phosphorylation was effectively blocked. In contrast, OKT4, which does not block



either HIV binding or infectivity^{13,14}, also does not block HIV-induced CD4 phosphorylation (Fig. 1c, lane 4). These data indicate that CD4 phosphorylation results from the specific interaction of CD4 and viral gp120. Furthermore, the ability of antibodies against CD4 (Leu3a but not OKT4) and gp120 (anti-gp120) to block CD4 phosphorylation correlates with their ability to inhibit viral infection¹³⁻¹⁷.

HIV infection is a stepwise process initiated by binding of HIV to CD4¹³⁻¹⁷. After binding, HIV enters the cell either by direct fusion of the virus envelope with the plasma membrane^{7,8} or by receptor-mediated endocytosis of a CD4/HIV complex¹⁸. Regardless of mechanism, viral entry is dependent on the presence of CD4 and occurs within 30–60 min of exposure to virus¹⁸. To determine the kinetics of HIV-induced CD4 phosphorylation, cells were incubated with HIV for up to two hours and the phosphorylation status of CD4 examined (Fig. 1d). HIV-induced phosphorylation occurred within 5 min, was maximal by 30 min and was maintained for up to 2 h in the presence of virus. The kinetics of CD4 phosphorylation correlate well with the kinetics of viral entry¹⁸, indicating that CD4 phosphorylation precedes or occurs concomitantly with viral entry.

The initial phosphorylation studies were performed on cloned CD4⁺ lymphocytes maintained in culture with alloantigen and interleukin-2¹⁹. These cells were used because they homogeneously express high levels of CD4 and respond to the same alloantigenic stimuli. We therefore reasoned that their response would approximate that of a single circulating CD4⁺ lymphocyte. To demonstrate that these phosphorylation responses are relevant to a mixed population of T-lymphocytes, we extended our phosphorylation studies to peripheral T-lymphocytes (Fig. 2). The basal level of CD4 phosphorylation in these cells is higher than that of CD4⁺ clones, possibly due

Fig. 1 HIV induces phosphorylation of its cellular receptor, CD4. **a**, ³²P-labelled CD4 was isolated from cells incubated with: 100 ng ml⁻¹ TPA (lane 1); 10 μ M H-7 followed by TPA (lane 2); HIV (100,000 c.p.m. reverse transcriptase activity per 1×10^6 cells) (lane 3); H-7 followed by HIV (lane 4); mock inoculum devoid of virus (lane 5); H-7 followed by mock inoculum (lane 6). **b**, Autoradiograph of [³²P]phosphopeptide maps of CD4 isolated from cells incubated with HIV or TPA. **c**, ³²P-labelled CD4 isolated from cells incubated with: HIV (lane 1); 10 μ g ml⁻¹ Leu3a (Becton Dickinson) followed by HIV (lane 2); HIV that had first been incubated with 10 μ g ml⁻¹ of anti-gp120 (DuPont-NEN) (lane 3); 10 μ g ml⁻¹ OKT4 (Ortho Diagnostics) followed by HIV (lane 4); the indicated antibody alone (lanes 5–7). **d**, ³²P-labelled CD4 isolated from cells treated with HIV for the indicated times. Results are plotted as the extent of CD4 phosphorylation (optical density) against time.

Methods. Cloned human CD4⁺ lymphocytes maintained in culture in the presence of IL-2 and alloantigen¹⁹ were washed twice with phosphate-free RPMI medium and equilibrated with [³²P]orthophosphate (Amersham, 100 μ Ci ml⁻¹) for 1 h at 37 °C. The cells (1×10^7 per point) were then incubated at 37 °C with the indicated agents. Cells were incubated with HIV or mock inoculum for 30 min or TPA for 10 min. H-7, Leu3a or OKT4 was added to cultures 30 min before addition of HIV or TPA. HIV was incubated with antibody to gp120 for 30 min before addition of the virus to cells. Following the incubations, cells were chilled to 4 °C, washed twice with ice-cold phosphate buffered saline (PBS) and lysed in PBS containing 1% (w/v) Nonidet P-40, 10 mM EGTA, 5 mM MgCl₂, 1 mM sodium vanadate, 20 mM sodium fluoride, 20 μ g ml⁻¹ leupeptin and 1 mM phenylmethylsulphonylfluoride (PMSF). All subsequent steps were performed at 4 °C. Soluble lysates were obtained following centrifugation at 8,500g for 5 min and incubated with 50 μ l Sansorbin (Calbiochem) for 30 min. The Sansorbin was removed by centrifugation and the lysates incubated with OKT4 (10 μ g IgG per 10^7 cells) for 30 min. Sansorbin (50 μ l; binding capacity of 2.11 mg IgG ml⁻¹) was added to the lysates and incubation continued for 30 min. The immunoprecipitates were washed five times with buffer containing PBS, 0.2% (w/v) Nonidet P-40, 1 mM sodium vanadate, 20 mM sodium fluoride, 2 mM EDTA, 1 mM PMSF and solubilized in SDS sample buffer containing 0.1% 2-mercaptoethanol for SDS-PAGE analysis²⁹ and autoradiography. Peptide mapping was performed as described by Cleveland *et al.*³⁰. Exposure times were the same for all samples in a panel.

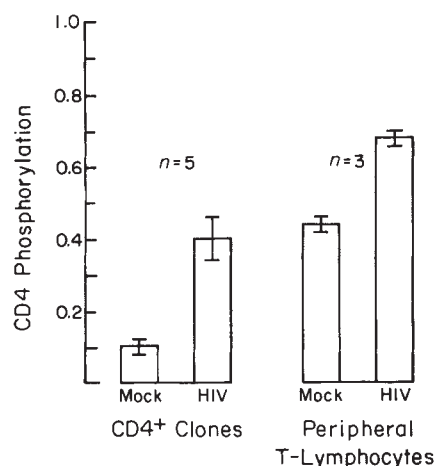


Fig. 2 Effect of HIV on CD4 phosphorylation in cloned CD4⁺ cells and peripheral T-lymphocytes. Cloned human CD4⁺ lymphocytes and peripheral T-lymphocytes²⁰ were equilibrated with [³²P]orthophosphate, incubated with HIV and CD4 was analysed by immunoprecipitation, SDS-PAGE and autoradiography as described in Fig. 1. The resulting CD4 bands were analysed densitometrically. The values represent the means of 5 (CD4⁺ clones) and 3 (peripheral T-lymphocytes) experiments $\pm$ s.e.m.

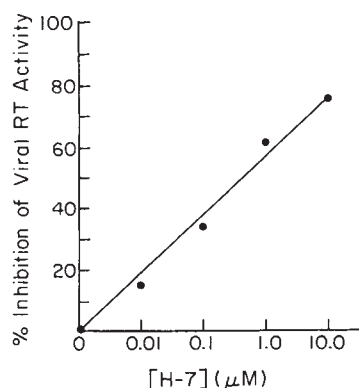


Fig. 3 Effect of various concentrations of H-7 on HIV infectivity. Data are presented as the per cent of inhibition of reverse transcriptase (RT) activity measured at 7 days following HIV infection in the presence of H-7.

Method. Peripheral T-lymphocytes were plated at 5×10^5 cells per ml in multiwell plates. The cells were treated with H-7 at the concentrations indicated for 30 min before infection. Infection was initiated by addition of HIV (10,000 c.p.m. RT activity per 1×10^5 cells) and cultures were maintained at 37°C for 7 days in the presence of the indicated concentration of H-7. H-7 was added only once at the start of the incubation period. RT activity was assayed from the culture supernatants as previously described³¹. Cell viability was $>95\%$, as determined by dye exclusion, in cells treated with up to $10 \mu\text{M}$ H-7 for 7 days in the absence of HIV.

to the presence of some activated T-lymphocytes within this population, as CD4 has been shown to be phosphorylated in activated lymphocytes¹⁰. But, as in CD4⁺ clones, treatment with HIV results in increased phosphorylation of CD4. Though it is possible that the phosphorylated CD4 may come from contaminating monocytes, this is unlikely as our isolated T-lymphocyte populations routinely contain $<1\%$ monocytes²⁰.

Next we investigated the mechanism by which H-7 inhibits HIV-induced CD4 phosphorylation. Fluorescence activated cell sorter analysis of H-7 treated cells demonstrated that H-7 had no effect on the surface expression of CD4 (data not shown), indicating that H-7 is not affecting HIV binding through modulation of CD4 surface expression. To test whether virus particles were capable of binding to H-7 treated cells, cells were subjected to thin-section electron microscopy and cell sections scored for cell surface-bound virus particles (Table 1). Cells treated with $10 \mu\text{M}$ H-7 before incubation with HIV had an average of 6.1 ± 0.6 virus particles per cell section, compared with an average of 2.1 ± 0.3 particles per cell section in cells treated with HIV alone. In contrast, cells incubated with HIV which had been treated with anti-gp120 antibody had only 0.2 ± 0.1 bound particles per cell section, demonstrating that this antibody interferes with the gp120-CD4 interaction(s) necessary for viral binding. These data indicate that H-7 does not inhibit HIV binding to CD4⁺ cells. Rather, H-7-treated cells appear to accumulate surface-bound HIV particles. Fluorescence microscopy confirmed the absence of viral protein within H-7 treated cells (data not shown). The simplest explanation for these findings is that H-7 interferes with the entry of cell surface-bound virus particles.

To determine the effect of H-7 on HIV infectivity, T-lymphocytes were incubated continuously with HIV in the presence of various concentrations of H-7 and infectivity assayed after seven days (Fig. 3). H-7 inhibits viral infectivity in a dose-dependent fashion, with a 50% reduction of infectivity seen at $\sim 1 \mu\text{M}$ and a 75% reduction at the highest concentration of H-7 used ($10 \mu\text{M}$). Interestingly, H-7 inhibits HIV infectivity at concentrations that effectively inhibit protein kinase C¹². Furthermore, $10 \mu\text{M}$ H-7 inhibits TPA-induced down-modulation of cell surface CD4 by 86% as determined by fluorescence activated cell sorter analysis (data not shown).

Accumulating evidence suggests that receptor phosphoryla-

Table 1 Effect of H-7 on HIV binding to CD4⁺ lymphocytes

Treatment	Number of virus particles bound/cell section
HIV	2.1 ± 0.3
HIV/H-7	6.1 ± 0.6
HIV- $\alpha\text{gp}120$	0.2 ± 0.1
Control	0

Cell sections were visualized by thin-section electron microscopy and scored for the number of cell surface-bound virus particles. Results represent the mean of cell-surface particles per cell section $\pm$ s.e.m. 20 cell sections were analysed in each treatment group. Cells and virus were treated as described in Fig. 1. After incubation, cells were washed twice in phosphate buffered saline (PBS) and fixed in PBS containing 1% paraformaldehyde and 1% glutaraldehyde for 30 min at 4°C . The cells were post-fixed with 1% OsO₄, dehydrated, embedded, sectioned and stained with uranyl acetate and lead citrate before viewing.

tion may be involved in the mechanism of receptor internalization²¹⁻²⁸. Our data indicate that HIV-induced phosphorylation of CD4 precedes or occurs concomitantly with viral entry. Blockade of HIV-induced CD4 phosphorylation with the protein kinase C inhibitor H-7 effectively inhibits viral infectivity in long-term cultures, possibly by interfering with the entry of cell surface-bound virus particles. Our results indicate that HIV-induced CD4 phosphorylation could be involved in the mechanism of viral entry. Furthermore, blockade of HIV-induced CD4 phosphorylation could represent a potential target for the development of effective antiviral therapy.

Note added in proof: A recent report demonstrates that the purified gp120 alone is apparently unable to induce CD4 phosphorylation in SupT1 cells³². We have also observed that a partially purified preparation of gp120 (gift of Drs Steven W. Pyle and Larry O. Arthur, Program Resources, Inc., NCI, FCRF) does not induce CD4 phosphorylation in peripheral T-lymphocytes. This suggests that binding of intact virus containing gp120 to CD4, instead of purified gp120 alone, may be required to mediate CD4 phosphorylation.

Received 21 December 1987; accepted 11 April 1988.

1. Fahey, J. L. *et al. Am. J. Med.* **76**, 95-100 (1984).
2. Barre-Sinoussi, F. *et al. Science* **220**, 868-871 (1983).
3. Gallo, R. *et al. Science* **224**, 500-503 (1984).
4. Levy, J. A. *et al. Science* **225**, 840-842 (1984).
5. Lifson, J. D., Reyes, G. R., McGrath, M. G., Stein, B. S. & Engleman, E. G. *Science* **232**, 1123-1127 (1986).
6. Lifson, J. D. *et al. Nature* **323**, 725-728 (1986).
7. Stein, B. S. *et al. Cell* **49**, 659-668 (1987).
8. Kowalski, M. *et al. Science* **237**, 1351-1355 (1987).
9. Hoxie, J. A., Matthews, D. M., Callahan, K. J., Cassel, D. L. & Cooper, R. A. *J. Immun.* **137**, 1194-1201 (1986).
10. Acres, R. B., Conlon, P. J., Mochizuki, D. Y. & Gallis, B. *J. Biol. Chem.* **261**, 16210-16214 (1986).
11. Weyand, C. M., Goronzy, J. & Fathman, C. G. *J. Immun.* **138**, 1351-1354 (1987).
12. Hidaka, H., Inagaki, M., Kawamoto, S. & Sasaki, Y. *Biochemistry* **5036-5042** (1984).
13. McDougal, J. S. *et al. J. Immun.* **135**, 3151-3162 (1985).
14. McDougal, J. S. *et al. Science* **231**, 382-385 (1986).
15. Sodroski, J., Goh, W. C., Rosen, C., Campbell, K. & Haseltine, W. A. *Nature* **322**, 470-474 (1986).
16. Dalglish, A. G. *et al. Nature* **312**, 763-766 (1984).
17. Klatzmann, D. *et al. Nature* **312**, 767-768 (1984).
18. Maddon, P. J. *et al. Cell* **47**, 333-348 (1986).
19. Sredni, B., Volkman, D., Schwartz, R. H. & Fauchi, A. S. in *Isolation, Characterization and Utilization of T-Lymphocyte Clones* (eds Fathman, C. G. & Fitch, F. W.) 440-452 (Academic, New York, 1982).
20. Julius, M. H., Simpson, E. & Herzenberg, L. A. *Eur. J. Immun.* **3**, 645-651 (1973).
21. Cochet, C., Gill, G. N., Meisenhelder, J., Cooper, J. A. & Hunter, T. *J. Biol. Chem.* **259**, 2553-2558 (1984).
22. Iwashita, S. & Fox, C. F. *J. Biol. Chem.* **259**, 2559-2567 (1984).
23. Davis, R. J. & Czech, M. P. *J. Biol. Chem.* **259**, 8545-8551 (1984).
24. Stadel, J. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **80**, 3173-3177 (1983).
25. Cantrell, D. A., Davies, A. A. & Crumpton, M. J. *Proc. natn. Acad. Sci. U.S.A.* **82**, 8158-8162 (1985).
26. May, W. S., Jacobs, S. & Cuatrecasas, P. *Proc. natn. Acad. Sci. U.S.A.* **81**, 2016-2020 (1984).
27. May, W. S., Sahyoun, N., Jacobs, S., Wolf, M. & Cuatrecasas, P. *J. Biol. Chem.* **260**, 9419-9426 (1985).
28. Klausner, R. D. *et al. Proc. natn. Acad. Sci. U.S.A.* **81**, 3005-3009 (1984).
29. Laemmli, U. K. *Nature* **227**, 680-685 (1970).
30. Cleveland, D. W., Fischer, S. G., Kirschner, M. W. & Laemmli, U. K. *J. Biol. Chem.* **252**, 1102-1106 (1977).
31. Adachi, A. *et al. J. Virol.* **59**, 284-292 (1986).
32. Hoxie, J. *et al. J. Immun.* **140**, 786-795 (1988).

Gold cluster-labelled antibodies

J.F. Hainfeld

The Fab' fragments of antibodies can be combined with eleven gold-atom clusters to produce the smallest gold-conjugated antibody probes yet developed.

COLLOIDAL gold bound through protein A to the Fc portion of antibodies is widely used as an electron-dense immunolabel by electron microscopists¹ (Fig. 1a). A 5-nm colloidal gold particle produces a probe with an overall dimension of 23 nm; 27.5-nm ferritin-antibody conjugates² are another popular label. While both have sufficient resolution for many studies, higher resolution immunological mapping requires a smaller probe. This has been achieved³ by covalently coupling a stable eleven gold-atom (undecagold) cluster (Fig. 2) with an 0.8-nm electron-dense core to a sulfhydryl group within the antibody hinge region of an Fab' fragment. The resulting probe is 5.0 nm in diameter (Fig. 1b) — five to ten times smaller than other antibody electron microscopy labels. Its small size enables subunit labelling of macro-molecular complexes and even, in some cases, the domain mapping of epitopes on single molecules.

The undecagold antibody probe also can mark sites inaccessible to the larger immunolabels, and infiltrates more quickly. A further advantage is its stability: once made, it can be kept for months, and the covalent gold-antibody bond allows it to remain conjugated over a wide pH or salt concentration range, conditions under which the non-covalently bound colloidal gold conjugates break down. The procedure used for preparation is mild, so that most antibodies (either polyclonal or monoclonal) can be labelled with undecagold. Two disadvantages are the inavailability of the gold clusters from commercial sources, and the poor visibility of the clusters in commercial electron microscopes. The clusters are clearly visualized in a high resolution scanning transmission electron microscope (STEM), however.

Probe composition

Undecagold clusters have been known for 19 years⁴, but have only more recently been made water-soluble⁵ for use in biological applications. The clusters consist of a dense core of eleven gold atoms with an organic shell of seven derivatized triphenylphosphines. In the form shown (Fig. 2), there are 21 primary amino groups on the periphery of the clusters, which may be used to attach covalently proteins or other molecules. The clusters are stable in the electron microscope⁶ and have been used to mark biotin binding sites on avidin⁷ to a 1-nm resolution — the first time isolated biomolecules were

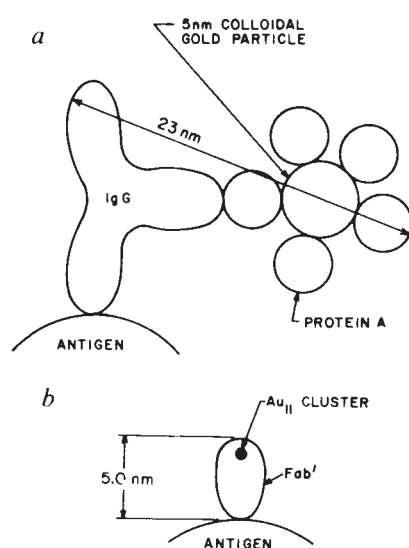


Fig. 1 Old and new antibody labels drawn for size comparison. *a*, A small conventional label formed by non-covalently adhering staphylococcal protein A to a colloidal gold particle and binding it to the Fc region of an IgG. (Overall dimension, 23 nm.) *b*, The Fab'-gold label, with a maximum dimension of 5.0 nm.

labelled to such a high spatial resolution for electron microscopy.

An important refinement in the use of the clusters was the synthesis of a monofunctional undecagold cluster with only one reactive group on its periphery, for labelling a specific molecular site. This was accomplished by both Safer *et al.*⁸ and Reardon and Frey⁹, independently.

Subsequently, a number of other proteins¹⁰, including actin¹¹, fibrinogen¹² and the carbohydrate moiety of the glycoprotein haptoglobin¹³, have been labelled with undecagold. (It should be pointed out that preparation of the undecagold clusters is not recommended for persons unfamiliar with anhydrous organic syntheses.)

The use of undecagold to label antibody fragments arose from its use with larger biomolecules. The strategy for the preparation of undecagold-antibody probes involves trimming the antibody down to a small active fragment, Fab', which conveniently has one or more free sulfhydryls available in the hinge region for reaction with a sulfhydryl-specific monofunctional form of the undecagold cluster. The gold cluster used in this work is that described by Safer⁸, where 20 of the 21 peripheral phosphine groups are nonreac-

tive and one is derivatized into a maleimide.

Making the probe

To prepare the probes, antibodies first are digested with pepsin to form F(ab')₂ fragments. Normally, rabbit polyclonals are incubated overnight at 37 °C with a 2 per cent by weight pepsin solution at pH 3.5 in sodium acetate. Mouse monoclonals require differing amounts of pepsin, pH and time to maximize the yield of F(ab')₂ fragment¹⁴. Because pepsin digests the Fc region into smaller fragments, a gel exclusion column can be used without a staphylococcal protein A binding step to purify the F(ab')₂ fragments. The fragments are next reduced to Fab' by treatment with 20 mM dithiothreitol (DTT) under N₂ for 2 hours. DTT is subsequently removed by gel exclusion chromatography. Meanwhile, the undecagold cluster, normally stored in methanol as a monoaminopropyl compound, is activated with N-methoxycarbonylmaleimide to add the maleimide group just before use.⁸ The Fab' and monomaleimidopropyl gold cluster are then mixed (using a 3–10-fold molar excess of gold cluster), and incubated overnight at 4 °C. Excess gold cluster (*M_r* = 5,000) is removed from the Fab' (*M_r* = 50,000) by a gel exclusion column. An example of a labelled monoclonal antibody is shown in Fig. 3.

New horizons

The procedure described is generally applicable to most antibodies, except

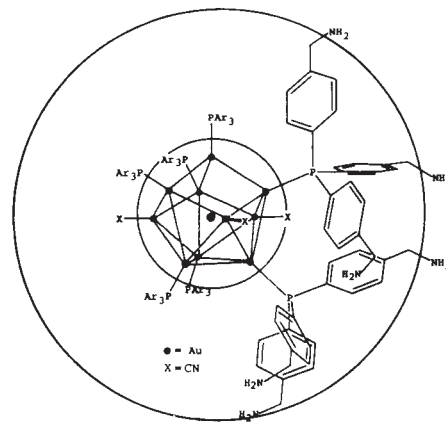


Fig. 2 Water-soluble undecagold cluster. The eleven-atom gold core is surrounded by seven derivatized triphenylphosphines, whose 21 amino groups may be used to attach other proteins.

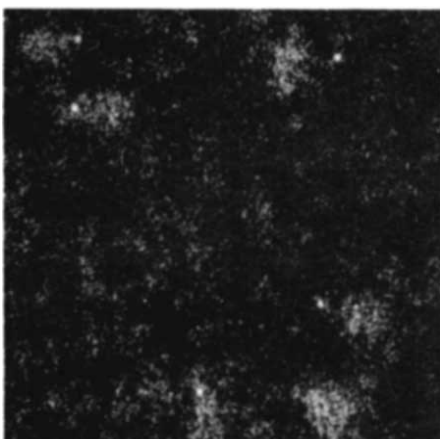


Fig. 3 STEM micrograph of monoclonal Fab' fragments bound to undecagold clusters. The Fab' fragments are $5 \times 3 \times 2$ nm in size; the gold clusters are visible as bright spots on one end of the Fab' fragments.

those few from which $F(ab')_2$ fragments cannot be readily prepared. It is possible to attach more gold clusters to antibodies containing more than one sulfhydryl group in the hinge region.

The future of this new antibody probe may not be limited to electron microscopy. Its unique properties of multiple atom labelling, small size, chemical stability and well-defined attachment location may suggest some interesting medical applications. Other clusters, such as tungstate, are being examined for conjugation in a similar fashion.¹⁵ Probes with larger clusters which are visible in conventional electron microscopes will widen the use of the technique, while the creation of smaller clusters and fragments will increase the resolution of immunolabels. The application of this new class of cluster immunoconjugates has just begun. □

James F. Hainfeld is at the Brookhaven National Laboratory, Biology Department, Upton, New York 11973, USA. For more information, fill in reader service number 100.

1. Faulk, W.P. & Taylor, G.M. *Immunochemistry* **8**, 1081-1083 (1971).
2. Singer, S.J. *Nature* **183**, 1523-1524 (1959).
3. Hainfeld, J.F. *Science* **236**, 450-453 (1987).
4. McPartlin, M., Mason, R. & Malatesta, L. *J. chem. Soc. chem. C*, **334**, (1969).
5. Bartlett, P.A., Bauer, B. & Singer, S.J. *Annls Am. chem. Soc.* **100** 5085-5089 (1978).
6. Wall, J.S., Hainfeld, J.F., Bartlett, P.A. & Singer, S.J. *Ultramicrosc.* **8**, 397-402 (1982).
7. Safer, D., Hainfeld, J.F., Wall, J.S. & Riordan, J. *Science* **218**, 290-291 (1982).
8. Safer, D., Bolinger, L. & Leigh, J.S. *J. inorg. Biochem.* **26**, 77-91 (1986).
9. Reardon, J.E. & Frey, P.A. *Biochemistry* **23**, 3849-3856 (1984).
10. Hainfeld, J.F. & Wall, J.S. in *Recent Advances in Electron and Light Optical Imaging in Biology and Medicine* Vol. 483 (ed. A.P. Somlyo) 181-187 (New York Academy of Sciences, New York, 1986).
11. Safer, D., Hainfeld, J., & Wall, J.S. *Biophys. J.* **47**, 128a (1985).
12. Mossesson, M.W. et al. in *Fibrinogen and Its Derivatives* (eds Muller-Berghaus, G. et al.), 3-15 (Elsevier, Amsterdam, 1986).
13. Lipka, J.J., Hainfeld, J.F. & Wall, J.S. *J. Ultrastruct. Res.* **84**, 120-129 (1983).
14. Parham, P. *J. Immunol.* **131**, 2895-2902 (1983).
15. Foley, C.J., Maellia, L.E., Hainfeld, J.F. & Wall, J.S. in *Proc. 45th Ann. Electron Microsc. Soc. Amer.* (ed. Bailey, G.W.) 760-761, (San Francisco Press, 1987).

Labelling techniques

The frontiers of molecular biology and immunology have been expanded through the use of radioisotopic and fluorescent labels. Several new reagents are described below.

BOEHRINGER Mannheim Biochemicals has recently announced the launch of its **nonradioactive DNA labelling and detection kit** (Reader Service No. 101). The kit can be used routinely to detect 0.1 pg of DNA in less than 16 hours without significant background, says the company — a degree of sensitivity which should allow the detection of single-copy genes in 1 µg of genomic mammalian DNA. The system can be used in dot-, slot- or Southern blots, colony and plaque hybridizations, and *in situ* hybridizations. The kit includes all components necessary for the labelling of DNA by random primed incorporation of digoxigenin-dUTP, and the detection



A non-radioactive DNA labelling kit from BMB.

of the labelled DNA with a specific anti-digoxigenin-alkaline phosphatase conjugate and subsequent colour reaction.

Amersham has a new **biotinylation kit** for the custom production of biotinylated antibodies using the spacer arm reagent incorporated into the current range of Amersham products (Reader Service No. 102). The kit allows researchers to prepare biotinylated primary antibodies and reduce the number of steps in immunocytochemical procedures while retaining the advantages of the biotin-streptavidin labelling system. The kit may also be used to biotinylate second antibodies where they are unavailable commercially, or to label enzymes and receptor ligands. Biotinylation reagent, buffer and purification columns for five reactions are included in the kit, along with a full protocol. Amersham says the spacer arm biotinylation reagent reduces steric hindrance between the biotin molecule and the macromolecule to which it is bound, yielding higher sensitivity.

The free 1988 **catalogue** from Dako Corporation, a division of Dakopatts, describes over 600 monoclonal and polyclonal antibodies, immunoenzymatic staining kits and related products (Reader Service No. 103). New products in the

catalogue include monoclonal antibodies to cytomegalovirus, Ki-1 (Reed-Sternberg), HLA-DR/alpha and myeloid/histiocyte antigen. A line of FITC-conjugated monoclonals is introduced in the catalogue, listing antibodies to pan T-cell, pan B-cell, CD8 and CD4 cells. More $F(ab')_2$ antibody fragments — some offered conjugated to horseradish peroxidase — have also been added.

BRL has a **photobiotin labelling system** which can be used as an alternative to nick translation for the nonradioactive labelling of DNA probes (Reader Service No. 104). By exposure to a strong light source, photobiotin links biotin groups to DNA. The system includes reagents for the preparation of DNA probes, control DNA to monitor performance, protocols for carrying out the procedure, and procedures for probe hybridization. BRL says one kit can label up to 50 µg of DNA.

MSD Isotopes offers a wide variety of **stable isotope labels** for biochemistry, chemistry and other disciplines. The MSD Isotopes catalogue lists deuterium, carbon-13, nitrogen-15, oxygen isotopes and labelled compounds (Reader Service No. 105). Stable isotopes may be used to define metabolic pathways, to follow organic reactions or in GC-MS for isotope internal standards. Specific products listed in the free catalogue include pollutants, drugs, biochemicals, spin labels, algal products and NMR solvents.

Zymed is offering **protein G** isolated from *Streptococcus* and **recombinant protein G** produced in a strain of *Escherichia coli* for use in ELISAs, antibody purification and other applications (Reader Service No. 106). Zymed also sells conjugates of recombinant protein G to biotin, FITC, Texas Red, alkaline phosphatase, horseradish peroxidase and Sepharose 4B. Zymed says its protein G reagents bind to all four subclasses of human IgG and to IgG from mouse, rabbit, guinea pig, rat, goat, sheep, cow and horse, and do not bind to human IgA or IgM. Its recombinant protein G will not bind human serum albumin, says the company. □

These notes are compiled by Carol Ezzell from information provided by the manufacturers. To obtain further details about these products, use the reader service card bound inside the journal. Prices quoted are sometimes nominal, and apply only within the country indicated.